PROJECTE FINAL DE CARRERA

MÉTODOS DE PROCESADO DE DATOS HIPERESPECTRALES APLICADOS AL ÁMBITO DE LA OCEANOGRAFÍA BIOLÓGICA

(Processing Methods of Hyperspectral Data Applied to Biological Oceanography)

ESTUDIS	·	Eng. de Telecomunicació
AUTOR	·	Sergio Pérez Lezcano
DIRECTOR	·	Ismael Fernández Aymerich
ANY	·	Septiembre de 2011

Resum del projecte

El fitoplancton és objecte de gran atenció en l'àmbit de la biologia marina. Aquests èssers microscòpics són els productors primaris dels oceans, a més de ser importants captors de CO_2. La seva caracterització proporciona informació molt valiosa sobre l'estat d'un ecosistema tan fràgil i important per l'èsser humà com és el marí. La necessitat d'explorar un medi tan inaccessible ha motivat el desenvolupament de tota una tecnología especialitzada, destacant la instrumentació òptica. Com a part del sistema dedicat a la fotosíntesi, les cèl·lules de fitoplancton tenen uns pigments capaços d'absorvir llum, part de la qual pot ser reemessa en forma de fluorescència. Aquesta energia lumínica pot ser induida i mesurada mitjançant sensors òptics hiperespectrals per obtenir un espectre complet. Com que cada espècie te una combinació de pigments particular, també ho serà la resposta en freqüència de la seva emissió lumínica. Aquest principi és el que s'util·litza en aquest projecte per intentar identificar diferents classes de fitoplancton a partir del seu espectre de fluorescència. La òptica, la estadística i l'aprenentatge de maquines s'util·litzen per fer possible en el futur un sistema automàtic per caracteritzar la columna d'aigüa.

Resumen del proyecto

El fitoplancton es objeto de gran atención en el ámbito de la biología marina. Estos seres microscópicos son los productores primarios de los océanos además de ser importantes captadores de CO_2. Su caracterización proporciona una valiosa información sobre el estado de un ecosistema tan frágil e importante para el ser humano como es el marino. La necesidad de explorar un medio tan inaccesible ha motivado el desarrollo de toda una tecnología especializada, entre la que se encuentra la instrumentación óptica. Como parte del sistema dedicado a la fotosíntesis las células de fitoplancton tienen unos pigmentos capaces de absorber luz, parte de la cual puede ser re-emitida en forma de fluorescencia. Esta energía lumínica puede ser inducida y medida mediante sensores ópticos hiperespectrales para obtener un espectro completo. Al tener cada especie una combinación de pigmentos particular también lo será la respuesta en frecuencia de su emisión lumínica. Es este principio el que se aprovecha en este proyecto para tratar de identificar diferentes clases de fitoplancton a partir de su espectro de fluorescencia. La óptica, la estadística y el aprendizaje de máquinas son utilizados para tratar de hacer posible en el futuro un sistema automático para la caracterización de la columna de agua.

Abstract

In the field of marine biology, phytoplankton is the subject of many studies. These microscopic algae are the primary producers of the oceans as well as being an important absorber of CO_2. Its characterization provides valuable information about the state of the marine ecosystem, so fragile and important for human being. The need to explore such a hostile environment has motivated the development of a complete specialized technology, included advanced optical instrumentation. As part of the system devoted to photosynthesis, phytoplankton cells contain pigments capable of absorbing light, part of which can be re-emitted in the form of fluorescence. This kind of luminescence can be induced and measured with hyperspectral optical sensors to obtain a complete spectrum. As each species has a particular combination of pigments, so will be the frequency response of the light emission. This is the principle that attempts to exploit this project to try to identify different classes of phytoplankton from its fluorescence spectrum. Optics, statistics and machine learning tools are used to try to make possible a future autonomous system for the characterization of the water column.

Índice general

1. Introducción

1.1 Contexto del Proyecto

El Centro Mediterréneo de Ciencias Marinas y Ambientales (CMIMA), centro adscrito al Consejo Superior de Investigaciones Científicas (CSIC), alberga numeroso grupos de investigación de muy diversos campos como la biología, la física, la geología o la ingeniería genética, cuyo punto de interés común es la investigación del mar. Algunos de estos grupos forman parte de la Unidad de Tecnología Marina (UTM) integrada en su mayor parte por técnicos e ingenieros. La UTM realiza actividades de apoyo logístico y técnico a buques oceanográficos y bases polares, así como actividades de investigación, cuyo objetivo es la innovación tecnológica en el ámbito de las ciencias marinas. Actualmente existen diferentes proyectos para desarrollar tanto nuevos sistemas de instrumentación como métodos de procesado de las señales obtenidas de los diferentes sensores.

La importancia del estudio del mar tiene varios motivos. Por un lado el hecho de que los océanos cubran el 71% de la Tierra la convierten en el mayor ecosistema sobre ella. El mar además es un gran regulador del clima. Un buen ejemplo es el continente europeo. La latitud a la que se encuentra le debería conferir un clima más frío, pero la acción de la corriente del Golfo desplazando grandes masas de agua caliente desde el golfo de México contribuye en atemperarla. De alguna u otra forma la vida del ser humano siempre ha estado ligada al océano. Aproximadamente la mitad de la población vive en zonas costeras o en zonas relativamente cercanas a ellas, sosteniendo la economía de gran parte de ellas.

Ahora el hombre mira al mar con distintos ojos. Con temor porque el alcance las consecuencias de los cambios que puedan producirse en su dinámica no son del todo comprendidos. Con tristeza porque la acción del hombre resumida en sobreexplotación pesquera y contaminación de sus aguas han roto como siempre el equilibrio. Con esperanza porque el mar también es energía y ya se aprovecha la fuerza de sus mareas, sus olas y su vida, mediante los cultivos de algas. Con curiosidad porque también esto ha movido sus actos e impulsado la ciencia.

Con este afán se ha realizado este proyecto, con la idea de que la ingeniería es fundamental para el desarrollo de la ciencia. La ciencia es la suma de pequeñas contribuciones que se remontan a miles de años atrás. Este proyecto no pretende ser más que una pequeña contribución más. Tan pequeña, quizás, como los seres vivos que lo protagonizan: el fitoplancton.

Pequeñas pero vitales. Estas microalgas forman la base para todos los niveles de la cadena trófica en el mar y son una fuente primaria de oxígeno atmosférico y de absorción de CO_2. Se estima que estos seres autótrofos realizan aproximadamente la mitad de la actividad fotosintética del planeta. El fitoplancton influye en la abundancia y diversidad de organismos marinos, determina el funcionamiento de los ecosistemas y establece un límite máximo a la pesca.

En algunos casos las poblaciones de fitoplancton pueden crecer localmente de forma masiva. Este fenómeno, conocido como floración algal (o en términos coloquiales marea roja debido a la coloración que en ocasiones tienen las aguas que lo sufren), supone en muchos casos un problema ya que las floraciones están asociadas al crecimiento de algas tóxicas, dando lugar a lo que se conoce como HAB (Harmful Algal Bloom) (Schofield et al. 1999). Los HAB pueden tener repercusiones tanto en aspectos de salud pública como comerciales en acuicultura. Moluscos como los mejillones se alimentan de fitoplancton que retienen cuando filtran el agua. Si el fitoplancton contiene toxinas, su acumulación puede llegar a niveles perjudiciales para la salud y los bivalvos no podrían ser consumidos (Anderson et al. 1999).

El análisis de las especies de fitoplancton y su abundancia es una tarea rutinaria para monitorizar el océano. La toma de muestras es una tarea costosa en recursos humanos y tiene un alcance limitado en el espacio y en el tiempo. Algunos de los procesos que se desean estudiar tienen una variabilidad mayor de la que se puede medir mediante este procedimiento. Un sistema automático que utilizara medios ópticos para estudiar la presencia de fitoplancton de forma eficiente proporcionaría una valiosa cantidad de datos que permitirían avanzar en el estudio de este ecosistema.

Gran parte del conocimiento que se posee actualmente de los procesos físicos y biológicos que se dan en los océanos, es debido al avance tecnológico que han experimentado campos tan importantes como la instrumentación, la computación, el procesado y las comunicaciones. La unión de estas disciplinas ha permitido la evolución de los estudios encaminados a la monitorización de las constantes vitales de nuestro planeta.

Las dos principales vías para obtener medidas e información del océano son por un lado la teledetección activa y pasiva proporcionada por satélites o aviones equipados con sensores al efecto, y por otro las medidas *in situ* (en el lugar) tomadas con la ayuda de equipos e instrumentos oceanográficos. La teledetección es fundamental para obtener datos ópticos a escala regional o global que permitan hacer un estudio general de los procesos que ocurren sobre la superficie y a bajas profundidades ópticas (García-Weil). Las medidas *in situ* son esenciales para la calibración y validación de los datos obtenidos a través de la teledetección así como para el desarrollo de algoritmos, además de ofrecer datos complementarios bajo la superficie con alta resolución temporal (Chang et al. 2006).

Un avance importante dentro del ámbito de la detección es la aparición de instrumentos hiperespectrales (espectro-radiometría de alta resolución). Hasta su llegada los datos eran obtenidos en bandas individuales (información de un pequeño intervalo del espectro) o en decenas de bandas (multiespectral, normalmente las más significativas para el estudio a realizar) de forma que en determinadas aplicaciones no era posible captar detalles esenciales para discriminar entre espectros similares. Actualmente con este tipo de sensores hiperespectrales prácticamente no existe limitación pudiendo adquirir el espectro completo con cientos o miles de bandas (Chang et al. 2004). La utilización de este tipo de sensores, gracias a su alta resolución (por ej. Sensor Toshiba con resolución de 0,02 nm en un rango espectral de 200-850 nm) nos permitirá la aplicación de nuevas técnicas analíticas dentro del campo de la espectroscopia.

De forma paralela a la mejora de los sensores existe una evolución de las estructuras que se utilizan para la observación y estudio de los océanos. Entre ellas podemos encontrar en primer lugar las fijas, como las boyas ancladas, los trípodes y las plataformas oceanográficas, por otro aquellas que cuentan con cierta capacidad para desplazarse como las boyas de deriva o las sondas, y por último aquellas con una mayor movilidad y flexibilidad como los barcos y los vehículos submarinos autónomos (AUV en sus siglas en inglés), entre los que se incluyen los planeadores submarinos (Glider) (Perry & Rudnick 2003). Este desarrollo se ha beneficiado de la reducción en coste, consumo y tamaño de los instrumentos, permitiendo incorporar una mayor carta de pago y su utilización en proyectos de menor presupuesto (Pons et al. 2007). Prueba de ello es el desarrollo de proyectos cuya finalidad es el despliegue de instrumentos para la toma de muestras de parámetros físicos, bio-ópticos o biogeoquímicos a nivel global.

En este contexto, la Unidad de Tecnología Marina está desarrollando un programa de investigación encaminado al diseño de nuevos dispositivos que permita obtener medidas

hiperespectrales para poder estimar la presencia de diferentes tipos de algas en la columna de agua. El primer objetivo que se ha planteado es la clasificación automática de fitoplancton presente en la columna de agua utilizando como datos de entrada los espectros asociados a diferentes propiedades ópticas. El procesado de estos datos deberá ir encaminado, muy probablemente, a la detección de patrones y diferenciación de fuentes para lograr caracterizar la columna de agua *in situ*. Es también importante destacar que la finalidad de este estudio es ser aplicado en un dispositivo para realizar medidas, procesado y toma de decisiones en tiempo real para, según las adquisiciones realizadas durante un perfil, actuar en consecuencia realizando un estudio más detallado de las zonas de interés, o incluso tomando muestras de dichas zonas.

Los diferentes tipos de fitoplancton presentan una serie de propiedades ópticas - como la absorción o la fluorescencia - cuyos valores cambian en función de la longitud de onda, dando lugar a firmas espectrales diferentes. Mediante diferentes instrumentos ópticos, se pueden utilizar las medidas ópticas para cuantificar la biomasa fitoplanctónica o discriminar entre diferentes grupos. En el caso de la fluorescencia, los pigmentos presentes en el fitoplancton son los que les confiere una firma espectral particular (Seppälä 2009). En la mayoría de ellos el pigmento predominante es la clorofila a, sin embargo, son los pigmentos secundarios los que posibilitan la distinción entre los diferentes grupos.

La investigación en la identificación del contenido pigmentario del fitoplancton haciendo uso del espectro de fluorescencia se remonta varias décadas atrás. Las primeras técnicas trataban de discriminar entre los cuatro grandes grupos en los que se puede clasificar al fitoplancton (Yentsch & Phinney 1985). Estos cuatro grupos tienen un contenido pigmentario bien diferenciado, con lo que es suficiente con medir el ratio entre el valor de fluorescencia en la longitud de onda de máxima respuesta de la clorofila a (presente en todas las especies) y a la de los pigmentos accesorios o secundarios. Basado en esta idea Cowles et al. (1993) lograron diferenciar entre las especies que contienen ficoeritrina y los que carecen de ella a través del espectro de emisión realizando medidas in situ a lo largo de la columna de agua.

Muchos de los estudios relacionados se basan en el espectro de excitación de fluorescencia que se obtiene excitando secuencialmente a varias frecuencias y midiendo la emisión, por lo general, a la longitud de onda de máxima respuesta de la clorofila a (Poryvkina et al. 2000, Xupeng et al. 2010a). La ventaja del espectro de excitación es que es más rico en matices que el espectro de emisión, facilitando la discriminación. La principal desventaja es que su adquisición es lenta aunque se cuente con instrumentación hiperespectral debido a que debe excitarse a varias

frecuencias de excitación en diferentes slots de tiempo para medir la respuesta independiente en cada una de ellas. El espectro de emisión en cambio se puede medir de una sola vez usando la instrumentación adecuada, luego su utilización es factible para aplicaciones en las que se desea realizar rápidas adquisiciones in situ.

En otras aproximaciones al problema se utilizó la matriz de excitación-emisión completa (disposición continua de varios espectros de emisión para un barrido de excitaciones) (Oldham et al. 1985, Zhang et al. 2006), en Xupeng et al. (2010b) se hace lo propio juntando varios espectros de emisión de la matriz para formar uno continuo que aproveche tanto la capacidad discriminativa del espectro de excitación como el de emisión. En Aymerich et al. (2009) se dieron ya los primeros pasos para una clasificación basada exclusivamente en espectros de emisión.

1.2 Objetivos

Este PFC se enmarca dentro de la temática de una tesis doctoral cuya temática versa sobre el análisis de sistemas de adquisición y procesado de espectros de fluorescencia en general.

El trabajo del proyecto se centra en la posibilidad de discriminar entre clases de fitoplancton con la utilización exclusiva del espectro de emisión de fluorescencia. El estudio se basará en los datos espectrales suministrados medidos sobre cultivos de fitoplancton con representantes de tres de los grandes grupos pigmentarios, algunos de los cuales comparten la misma clase.

El principal objetivo es demostrar que el espectro de emisión puede proporcionar suficiente discriminación incluso entre especies con similar contenido pigmentario cuando éstas se encuentran en cultivos puros, es decir, midiendo sobre una especie a la vez.

Para lograr esto se pretenden aplicar diversas técnicas que ayuden a mejorar el procesado de datos hiperespectrales. Estas técnicas se integran dentro de un flujo de trabajo (Figura 1.1) cuyo objetivo final es la clasificación exitosa de las especies.

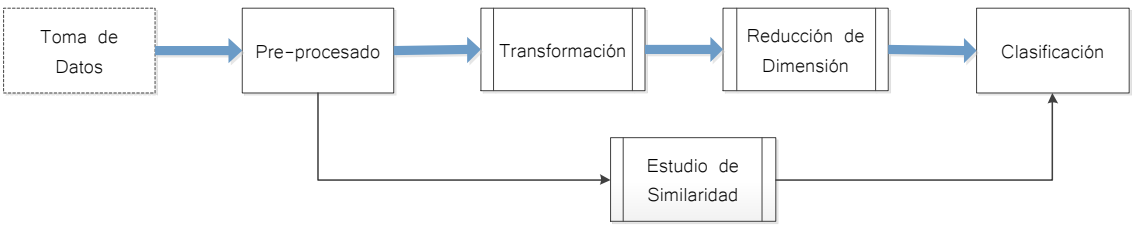


Figura 3.1. Diagrama de flujo.

En general los resultados de este proyecto pretenden motivar nuevas investigaciones que sigan los pasos necesarios para integrar este tipo de técnicas en dispositivos automáticos de detección in situ.

1.3 Estructura de la memoria

La presente memoria de proyecto cuenta con cinco capítulos principales, sin contar con la introducción y las conclusiones finales. En el capítulo dos se explican brevemente algunos conceptos sobre la fluorescencia, fenómeno en el que se basan los espectros disponibles cuya descripción preliminar se realiza al final del capítulo. En el tercero se analizan en profundidad los datos para aplicar las técnicas de pre-procesado que proceden para mejorar la calidad de los datos a la par que conservando toda la información en ellas disponible. En el cuarto se presentan algunas transformaciones aplicadas a los datos como parte del pre-procesado o como forma de encontrar nuevas formas de representación que mejoren la capacidad discriminativa. Es ésta capacidad la que se trata de evaluar en el siguiente capítulo, el usando, usando parámetros utilizados en el campo de la detección remota para ayudar a seleccionar el tipo de medida de distancia y clasificador más adecuado. En el sexto se ensayan algunos clasificadores y se presentan los resultados finales. El séptimo capítulo está dedicado a las conclusiones del proyecto y a señalar algunas posibles vías sobre las que deben de continuar el trabajo y las futuras investigaciones. Las referencias y la bibliografía adicional cierran la memoria.

2. Información hiperespectral de especies de fitoplancton

2.1 Introducción

El estudio del mar mediante instrumentación óptica permite tomar muestras no invasivas que pueden ser analizadas de forma rápida. Muchas técnicas de este tipo son ya aplicadas con éxito. Algunas de ellas se basan en medidas pasivas, que aprovechan la interacción de la luz ambiental con el agua de mar y los seres que la habitan, en su mayor parte microscópicos, que la habitan. Un inconveniente de estas técnicas es que su utilización está limitada al rango de profundidades en las que penetra suficiente luz solar. Los instrumentos para realizar medidas activas en cambio no necesitan la luz externa puesto que son ellos mismos la fuente de la excitación.

La fluorescencia, el fenómeno óptico sobre el que se basan los datos disponibles y por tanto sobre el que se basa el proyecto es un tipo de medida activa. Los principios básicos de su generación y el motivo por el que el fitoplancton es capaz de producirla son asuntos tratados en esta sección.

2.2 Fluorescencia

La utilización de la fluorescencia como herramienta de estudio ha sufrido un importante crecimiento en las últimas décadas y es hoy una metodología dominante en numerosos campos como la biotecnología, la medicina, la ingeniería genética o de utilización en técnicas como la citometría de flujo.

La fluorescencia es el proceso mediante el cual una sustancia absorbe luz y la re-emite a una longitud de onda distinta. Es un tipo de luminiscencia que en general consiste en la capacidad de absorber y re-emitir luz, lo cual también englobaría a la fosforescencia. La diferencia estribe en que la fluorescencia tiene lugar en una escala de tiempo mucho menor. Ciertas moléculas llamadas fluoróforos tienen la capacidad de ser excitadas, a través de la absorción de energía lumínica, a un estado excitado, el cual no puede mantener durante largo tiempo. El proceso se describe en la figura 2.1.

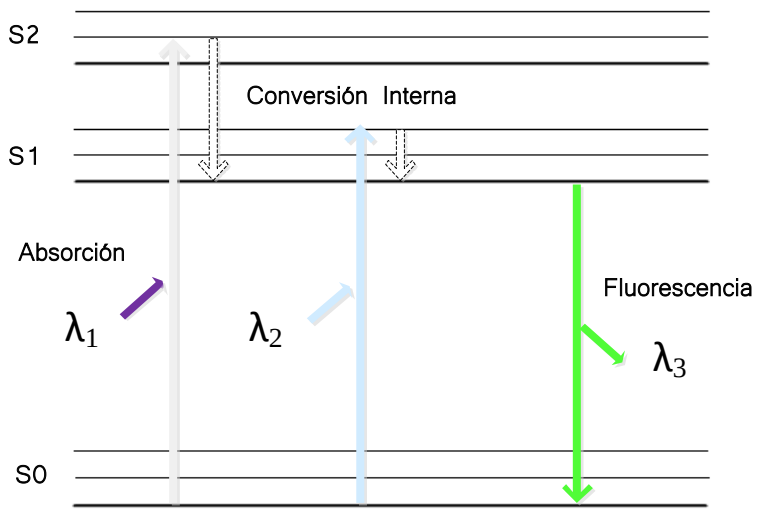


Figura 2.1. Representación del diagrama de Jablonski, en el cual se muestran los estados electrónicos de la molécula y las transiciones entre ellos.

Un fluoróforo inicialmente en un estado de baja energía o fundamental, absorbe un fotón mediante el cual pasa a un estado de energía superior. Rápidamente parte de la energía absorbida se pierde en un proceso llamado conversión interna, transformándose por ejemplo en calor, quedando la molécula en su estado energético más bajo. La energía restante es emitida en forma de luz, volviendo la molécula a su estado de baja energía. Dado que la frecuencia del fotón depende de su energía (Ecuación 2.1) al conservar menor energía que la absorbida en el momento de ser re-emitida, la frecuencia de emisión será menor que la de absorción, es decir, tendrá una longitud de onda (λ) mayor. El desplazamiento en frecuencia entre la onda de excitación y la de emisión recibe el nombre de desplazamiento de Stokes y es una característica del fluoróforo (Labowicz, p. 5).

$$E = \frac{hc}{\lambda} \qquad (2.1)$$

Una propiedad importante es que por lo general la longitud de onda de emisión no depende de la de excitación, propiedad que recibe el nombre de regla de Kasha (Labowicz, p. 7). Como se ilustra en la Figura 2.1, a pesar de que la molécula absorbe dos fotones que llevan a ésta a dos estados energéticos distintos, la conversión interna ocurre antes de que el fotón tenga tiempo de ser re-emitido, por lo que la energía disponible será la misma. Como al retornar al estado

fundamental el fotón puede terminar en estados de vibración distintos, la frecuencia emitida no es siempre exactamente la misma, variando en torno a una longitud de onda central con una cierta distribución, conformando el espectro de emisión.

La absorción de fotones por parte del fluoróforo no es igualmente efectiva para todas las longitudes de onda. Esto significa que la λ de emisión depende de la λ de excitación. Esto es utilizado para caracterizar materiales, sustancias o formas de vida como el fitoplancton. Si progresivamente excitamos una sustancia a diferentes longitudes de onda y medimos la cantidad de luz emitida (preferiblemente a la λ de máxima respuesta) podemos realizar una representación de la fluorescencia en función de la λ de excitación, llamado espectro de excitación. La forma de este espectro depende del contenido en fluoróforos de la sustancia en particular y permite diferenciarla de otras sustancias.

Otra forma de representar la fluorescencia es mediante su espectro de emisión, es decir la distribución de la energía lumínica emitida. En este caso estaríamos excitando a una sola λ y midiendo en una banda.

2.3 Propiedades ópticas del fitoplancton

El fitoplancton contiene pigmentos que realizan distintas funciones. La principal es la absorción de luz para realizar la fotosíntesis, transformación de la energía lumínica en energía química. Algunos de estos pigmentos también son fluoróforos y por tanto son capaces de emitir fluorescencia. No toda la energía que se absorbe es transformada en energía química. La emisión de fluorescencia es una de las maneras mediante la cual las células liberan el exceso de energía absorbida. Aproximadamente el 18% de la luz absorbida es utilizada durante las reacciones fotoquímicas, entre un 1 y un 3% se vuelve a emitir en forma de fluorescencia, y el resto es disipada en forma de calor (Kirk 1994).

Según el contenido en pigmentos, se clasifica a las algas y al fitoplancton en grandes grupos pigmentarios, por ejemplo las algas verdes, marrones o rojas en alusión a la coloración que las caracteriza. El principal pigmento presente en cualquier célula de fitoplancton es la clorofila a. Sin embargo es la presencia de pigmentos accesorios y sus proporciones las que son particulares de cada clase o especie. Como cada uno tiene su máximo de absorción en diferentes longitudes de onda, la forma del espectro de excitación varía para cada uno de ellos y permite distinguirlos.

El espectro de emisión también es característico de cada especie de fitoplancton, pero suele ser más uniforme y contener menos matices. Es por ello que ha recibido menor atención que el espectro de excitación para la caracterización de las especies. Su menor variación se debe a que muchos pigmentos no emiten fluorescencia directamente, sino a través de la clorofila a (Xupeng et al. 2010b). Éste fluoróforo tiene su máxima emisión alrededor de 680 nm (Tabla 2.1). Otros dos fluoróforos con emisión propia son la ficocianina, con emisión alrededor de 640, y la ficoeritrina que hace lo propio en torno a 570 nm.

Pigmento	Banda de emisión máxima
Chl a	680
Chl b	680
Chl c	680
Carotenos	680
Ficoeritrina	550-580
Ficocianina	630-660

Tabla 2.1. Bandas de emisión de fluorescencia máxima de los principales pigmentos.

2.4 Datos de laboratorio

Los datos experimentales que han sido proporcionados para la realización de este proyecto fueron tomados por un espectrofluorímetro Aminco Bowman Series 2. Éstos son el resultado de medir el espectro de excitación y emisión sobre una selección de cultivos de fitoplancton pertenecientes a diferentes grupos taxonómicos (Tabla 2.2). Fue realizada una medida diaria de cada cultivo hasta que las células finalmente morían. Por este motivo el número de muestras disponibles de cada especie es variable.

Especie	Clase	Grupo Pigmentario	Abreviación
Thalassiosira weissflogii	Bacillariophyceae	IV	Thwi
Dunaliella primolecta	Chlorophyceae	III	Duna
Pleurochrysis elongata	Primnesiophyceae	IV	Pl
Alexandrium minutum	Dinophyceae	IV	Amin
Isochrysis Galbana	Primnesiophyceae	IV	Iso
Ostreococcus sp.	Prasinophyceae	III	Ost

Tabla 2.2. Especies de fitoplancton de los datos disponibles.

El espectrofluorímetro utiliza una fuente de xenón de alta intensidad con espectro de emisión idealmente plano. La luz se transmite a través de un monocromador que selecciona la estrecha banda de excitación que va a ser utilizada. Dentro de compartimento de la muestra un obturador controla el tiempo de iluminación. Una muestra de referencia es tomada antes de polarizar e iluminar la sustancia bajo estudio. La señal de fluorescencia se mide en una dirección a 90º respecto de la de iluminación. Otro monocromador selecciona la banda de emisión deseada antes de incidir la luz sobre un fotomultiplicador para finalmente hacer un tratamiento digital.

Hubo dos grupos de cultivos independientes. En el primero de ellos la toma de datos se realizó barriendo las longitudes de onda de excitación desde 200 hasta 600 nm en pasos de 10 nm. Para cada λ de excitación se midió en el rango de longitudes de onda de emisión de 200 a 800 nm en pasos de 1 nm, para conferirle una resolución hiperespectral. Juntando los espectros de excitación y emisión disponibles se conforma la llamada matriz de excitación-emisión (EEM) (Figura 2.2). La segunda toma de datos se concentró en cuatro λs de excitación, una a 470 nm y otra a 490 nm, cuya emisión fue medida en el rango de 300 a 800 nm en pasos de 1 nm. Por tanto, de esta toma de datos se cuenta con un menor número de espectros de emisión.

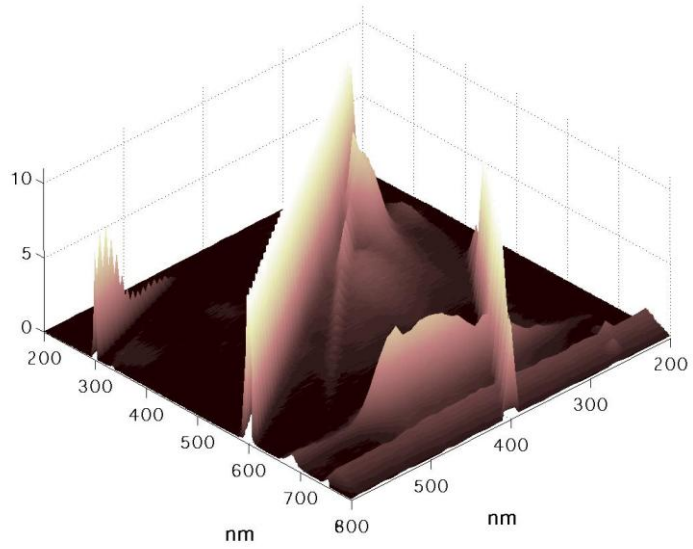


Figura 2.2. Matriz de excitación emisión (EEM) de una muestra de Duna.

En la figura 2.2 se observa la emisión de fluorescencia a través de la clorofila en torno a 680 nm. Esta es la zona de fluorescencia de emisión de interés de las especies bajo estudio. Como se aprecia, la fluorescencia no es el único fenómeno medido con el espectrofluorímetro. El efecto que presenta mayor magnitud es la dispersión de Rayleigh, cuyo primer orden se mide a la misma longitud de onda que la de excitación. Al enfocar luz sobre las muestras de agua de fitoplancton, al incidir sobre las células la luz se dispersa en varias direcciones en función de la forma y el tamaño que tenga, conservando la energía y por tanto la frecuencia por lo que es un tipo de dispersión elástica. Una de estas direcciones es la del receptor óptico y por tanto es recibida por el espectrofluorímetro. La dispersión de Rayleigh de segundo orden aparece también con una amplitud importante y posee una longitud de onda doble a la de excitación. Ésta atraviesa la zona de fluorescencia de interés para las longitudes de onda de excitación alrededor de 300-350 nm. También es visible la dispersión de Rayleigh de tercer orden, pero ya con mucha menor amplitud y a longitudes de onda muy bajas.

Otro efecto notable es la Generación de Segundo Harmónico (Second Harmonic Generation - SHG) (Franken et al.) que se manifiesta a la mitad de la longitud de onda de excitación. Este fenómeno se produce cuando dos fotones interaccionan de forma efectiva para formar uno nuevo con el doble de energía, y por tanto el doble de frecuencia. Afortunadamente la fluorescencia se produce en frecuencias menores a la de excitación por lo que esta zona de la matriz de EEM no es de interés. Sin embargo este efecto sí se hace notar en un rango de longitudes de onda comunes a la fluorescencia a través de la componente suma entre ella y la dispersión de Rayleigh de primer orden. Por ejemplo si se excita a 490 nm, la SHG se da a 245 nm y la componente suma a 735 nm. Dado que es de pequeña magnitud no se aprecia en la EEM de la figura 2.2, pero puede producir distorsión en las zonas del espectro de emisión de mayor longitud de onda.

Al igual que la de Rayleigh, la dispersión de Raman es debida a la interacción de la luz con las moléculas. Sin embargo la de Raman es una dispersión inelástica y tiene un desplazamiento a frecuencias menores que la de Rayleigh para el tipo Stokes. La dispersión de Raman comienza muy pegada a la de Rayleigh, pero a medida que aumenta la λ de excitación se va separando de ella. Puede producir distorsión en el espectro de fluorescencia por el lado de las longitudes de onda más bajas. No es visible en la figura 2.2 la dispersión de Raman de segundo orden pero aparece para longitudes de onda doble a la del principal.

También hay un efecto que se midió entre 730 y 800 nm independientemente de la λ de excitación. Su origen no ha sido precisado pero dado que aparece a frecuencias del infrarrojo puede tener un origen térmico.

Algunos de estos efectos pueden contribuir en la diferenciación entre especies. Por ejemplo la cantidad de dispersión de Rayleigh recibida depende de la concentración de células, de su tamaño y forma. Sin embargo el estudio de este proyecto se centra solo en la fluorescencia de emisión.

Si se analizan las EEM de un mismo cultivo durante el transcurso de las medidas diarias se observa como inicialmente suelen tener una baja respuesta de fluorescencia para progresivamente ir aumentando hasta un cierto nivel a partir del cual parece estancarse. Esto se debe a la curva de crecimiento del fitoplancton. Inicialmente hay pocas células y muchos nutrientes por lo que éstas se dividen y aumentan su concentración de forma exponencial hasta que se estanca por la escasez de nutrientes y finalmente mueren. Esto introduce una importante variabilidad de los espectros que tendrá que ser considerado.

3. Pre-procesado de información hiperespectral

3.1 Introducción

Antes de poder utilizar los espectros obtenidos con el espectrofluorímetro, éstos deben ser tratados adecuadamente de forma que se pueda eliminar todo aquello que se considera superfluo o que enmascara la información que se desea utilizar para caracterizar las especies. Dentro de nuestro diagrama de flujo, este paso seguiría a la adquisición de las muestras y precedería a cualquier tipo de transformación u extracción de información (figura 3.1).

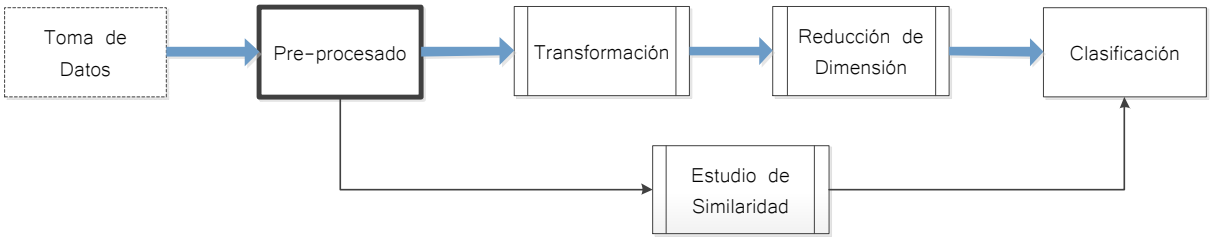


Figura 3.1. Diagrama de flujo.

En primer lugar se procedió a la selección del rango de longitudes de onda (λ) que se tomarían de cada espectro de fluorescencia, con el criterio de abarcar la mayor información con el menor número de bandas posible.

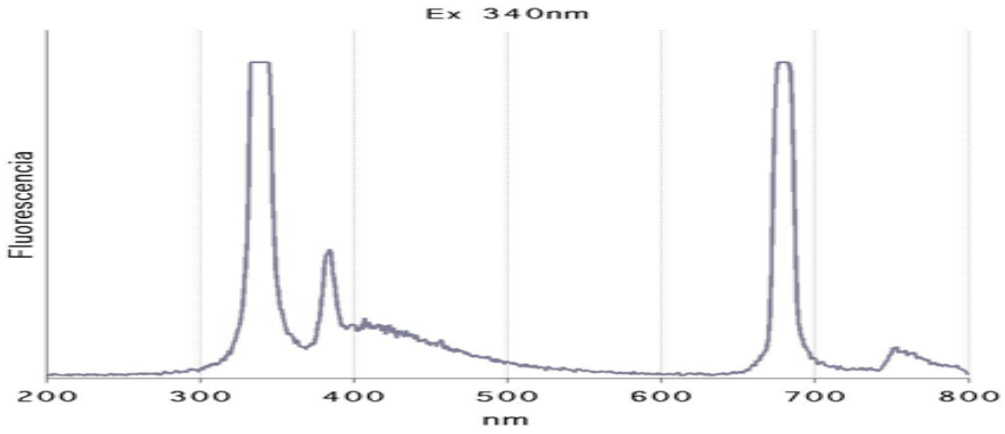
En la siguiente fase, la limpieza de los datos, se eliminan del conjunto de datos aquellas muestras que no se consideren representativas de su clase y que por tanto no sean buenos ejemplos en los cuales basarse para modelar la especie.

Cada longitud de onda de medida contiene el valor real de fluorescencia más una componente aleatoria debida a la suma de todas las fuentes de ruido que afectan al sistema de toma de muestras. El filtrado o suavizado de las curvas trata de atenuar su efecto de forma que aumente la relación señal a ruido en las mismas. Esta acción cobra mayor importancia si se pretende realizar un estudio a partir de la derivada de los espectros.

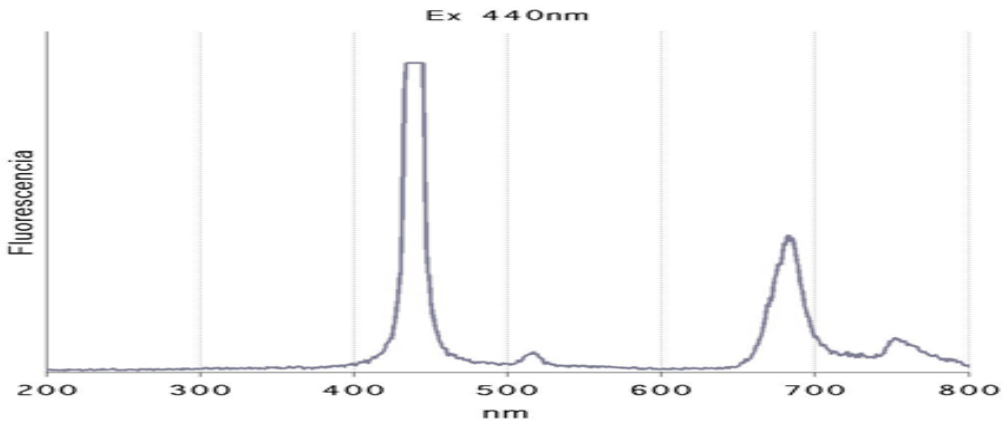
La concentración de células dentro del volumen del agua analizado determina la intensidad de fluorescencia que es posible medir. Para comparar curvas de fluorescencia obtenidas en distintos estadios de crecimiento de las poblaciones de fitoplancton, éstas deben ser normalizadas bajo un cierto criterio. Será éste, por tanto, la última fase del pre-procesado.

3.2 Selección de la banda de trabajo

Recordemos del apartado 2.4 que el conjunto de datos con el que se cuenta consiste en dos tomas de muestras independientes. En ambas se realizó un cultivo por cada especie a estudiar y se realizaron medidas diarias, pero en la primera de ellas éstas consistieron en una matriz EEM completa (Ex: 200 – 600 nm, Em: 200 – 800 nm) mientras que en la segunda se obtuvieron dos espectro de emisión (Ex: 470 nm y 490 nm) y un espectro de excitación (Em: 690 nm). Como ya se comentó, el trabajo se centro en los espectros de emisión. Por tanto, en principio se cuenta con 41 espectros de emisión por cada muestra perteneciente a la primera toma y 2 adicionales por cada muestra de la segunda.



(a)



(b)

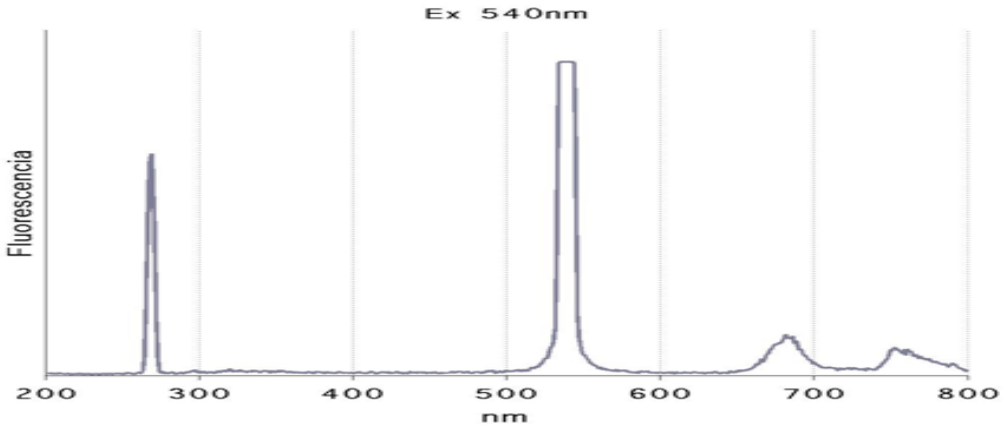


Figura 3.2. Espectros de emisión completos de la misma EEM de una muestra de PI para tres excitaciones distintas: 330nm (a), 430nm (b) y 530nm (c)

Para la primera selección de la banda de trabajo se usó un criterio puramente observacional. La figura 3.2 muestra tres ejemplos de espectros de emisión a distinta λ de excitación. Interesa tomar un rango que cubra la zona en la que se produce la fluorescencia, con máxima intensidad alrededor de 670–700 nm, prestando atención de que los extremos no se vean afectados por otros fenómenos ajenos a ella. En longitudes de onda de emisión por encima de 740 nm está presente en la mayor parte de las muestras una emisión independiente de la excitación, posiblemente de origen térmico, por lo que este será el límite superior de la banda.

Por el otro, como se aprecia en la figura 3.2, los scattering de rayleigh y de raman se encuentran más cerca de la zona de fluorescencia cuanto mayor es la λ de excitación. El scattering de raman es de menor intensidad y es casi despreciable para excitaciones por encima de 500 nm. El scattering de Rayleigh de primer orden es bastante notorio y su mayor longitud de onda de emisión es de 600 nm. Para evitar ambos efectos no se utilizarán las longitudes onda por debajo de 620 nm.

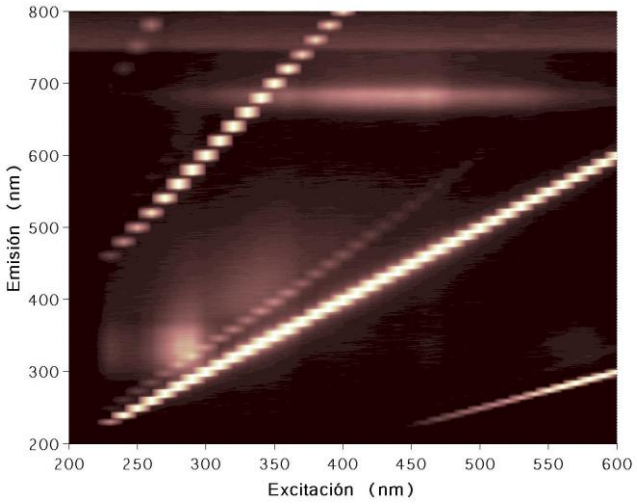


Figura 3.3. Visión XY de una matriz de excitación emisión (EEM) para una muestra de PI.

Sin embargo, el rango de trabajo habitual será de 630 a 730 nm, el cual abarca todo el espectro de fluorescencia de emisión de las especies bajo estudio. Con una resolución de 1 nm, se obtienen muestras de dimensión 101.

3.3 Limpieza de los datos

Cuando se cuenta con un set de muestras para clasificar, uno de los pasos previos consiste en realizar un pre-análisis para encontrar aquellas que, por algún motivo, se alejan de lo que es norma habitual en una determinada clase. En el ámbito de la estadística, a este tipo de muestras se les denomina valores atípicos o *outlier*, su traducción al inglés. En nuestro caso los *outliers* consisten en espectros que toman valores numéricos distantes de los del resto de la misma especie, ya sea en todas las bandas o en un subconjunto de ellas.

Una forma de detectar estas muestras atípicas es aplicar alguna técnica de agrupamiento o clustering sobre las muestras de una misma clase y tratar de identificar grupos intraclase. Esta tarea es más sencilla si se utiliza alguna proyección de los datos de dimensión n a otro más adecuado para su representación como dos dimensiones en el plano, o dos en el espacio. Si cada muestra es representada en el espacio mediante un punto, es fácil detectar aquellos que se encuentran aislados del resto. Un examen posterior a la muestra en cuestión puede esclarecer la causa de su diferenciación.

Otra manera es la inspección directa de los datos. Evidentemente, esta forma de proceder está reservada para aquellos casos en los que el volumen de los datos y el número de variables sea reducido, o en el caso de que la dimensión no sea baja, que al menos puedan ser representados fácilmente. Por último también es posible predecir la existencia de posibles outliers conociendo los factores que pueden tener influencia en la toma de medidas considerando el procedimiento que se llevó a cabo.

En el caso que nos ocupa, algunos outliers fueron detectados durante la aplicación de algunas técnicas que se verán en apartados posteriores, sin embargo en gran medida la inspección directa y la predicción han sido posibles dado que el número de muestras no es demasiado elevado y aunque la dimensión es relativamente alta, es sencillo representar o bien las muestras individuales, o las matrices de EEM completas.

Si se utiliza un set de datos con outliers para, por ejemplo, entrenar el modelo de un clasificador, su capacidad de generalizar para poder identificar correctamente nuevos datos puede verse afectada. Es preferible clasificar incorrectamente unas pocas muestras 'diferentes' que aumentar el error que se comete con un porcentaje alto de la población.

En cualquier caso, es importante tener constancia por lo menos de su existencia para no sacar conclusiones erróneas sobre alguna técnica o los datos. En otra fase del estudio pueden ser útiles para estudiar la robustez del método empleado. Por ejemplo se puede realizar un test con muestras de fluorescencia de baja intensidad para estudiar cual es la mínima necesaria para obtener un error de identificación máximo determinado.

Así pues, en el apartado anterior hemos considerado algunos de los efectos presentes en la matriz EEM para poner límite superior e inferior a la banda de los espectros de emisión. Sin embargo como deja patente la figura 3.3, hay otros efectos que pueden hacer variar las curvas de fluorescencia medidas y que aun no se han considerados.

El scattering de Rayleigh de segundo orden atraviesa por completo la zona de fluorescencia cuando el valor de la excitación es la mitad de las longitudes de onda de emisión en las que hay fluorescencia (alrededor de 690 nm). Así la máxima coincidencia entre la fluorescencia y el scattering se produce para una excitación de 350 nm, (recordemos que la excitación se realizó a saltos de 10 nm), como pone de manifiesto la figura 3.2a, pero afecta a la banda seleccionada para excitaciones desde 300 hasta 380 nm.

El scattering de Raman de segundo orden, aunque de una forma más sutil, afecta en la banda de 280–300 nm (Figura 3.4a), mientras que la frecuencia suma entre el scattering de Rayleigh y la Generación de Segundo Armónico (SHG), también de poca intensidad, afecta de forma desigual según sea la especie iluminada. Como ya se comentó, la intensidad del scattering depende del tamaño de las células de la especie y de su concentración, es decir, de su distribución por tamaño. Por ejemplo, este fenómeno óptico no lineal afecta en las medidas realizadas sobre *Thwi* y *Amin* dentro de la banda de trabajo seleccionada en el intervalo de excitación 470–490 nm (Figura 3.4b). Por debajo de estas longitudes de onda también tiene presencia pero se va extinguiendo progresivamente además de que el nivel de fluorescencia es suficientemente intenso como para que la relación señal a interferencia sea adecuada. Las especies *Duna* y *Pl* al provocar menos scattering se ven afectadas en menor medida, siendo perceptible su efecto para excitaciones a 480 y 490 nm. En *Iso* se consideró su influencia despreciable porque por un lado el scattering es reducido y por otro porque la fluorescencia recibida de esta especie es en general muy elevada.

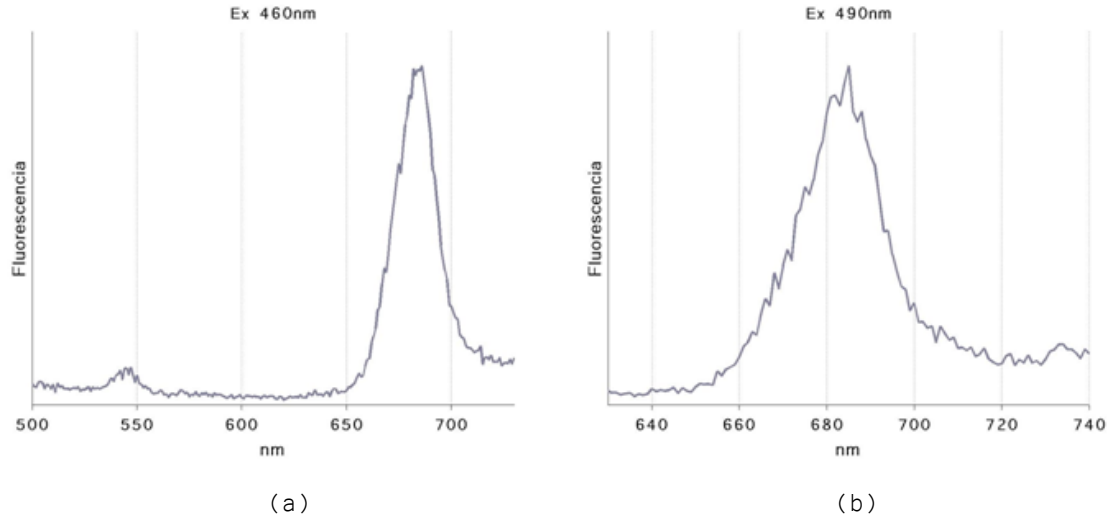


Figura 3.4. Espectros de emisión afectados por la dispersión de raman (a) y la componente suma de Rayleigh y SHG (b). La dispersión de Raman en este caso aparece alrededor de 490 nm. La componente suma se manifiesta como una pequeña protuberancia a unos 735 nm.

El último aspecto a tener en cuenta es la intensidad de fluorescencia mínima exigida a las muestras. Cuanto más intenso sea el espectro de emisión de fluorescencia, mejor definida estará

la curva y es de esperar que un clasificador tendrá menos problemas para etiquetarla. Por otro lado si se desea entrenarlo para que sea capaz de identificar señales de fluorescencia lo más débiles posible, es deseable proporcionarle señales de este tipo. En cualquier caso, simplemente estableciendo un umbral inferior para la intensidad máxima de fluorescencia se controlará el nivel mínimo de las curvas.

3.4 Suavizado de las curvas

Las curvas de fluorescencia que proporciona el espectrofluorímetro están afectadas en gran medida por el ruido. Las principales fuentes de ruido son la debida a la naturaleza aleatoria de los fenómenos lumínicos, el ruido térmico asociado a la circuitería y el ruido instrumental.

Un receptor óptico genera fotoportadores cuando sobre él inciden fotones. La probabilidad de que la llegada de un fotón genere un fotoportador sigue una distribución de Poisson, siendo por tanto una variable aleatoria. Las fluctuaciones que esto provoca en la fotocorriente es el denominado ruido de disparo. Otra fuente de ruido de los componentes ópticos es el ruido de oscuridad, o la respuesta del receptor en ausencia de estímulo luminoso. La utilización de un fotomultiplicador de buenas prestaciones reduce considerablemente estos fenómenos.

El ruido térmico es debido a la agitación térmica de los electrones en la circuitería electrónica por estar sometida a una cierta temperatura (por encima del cero absoluto). A diferencia del ruido de disparo, también está presente en ausencia de señal.

Varios factores intrínsecos al instrumento de medida utilizado, en este caso un espectrofluorímetro, pueden afectar a la medida. La no idealidad de los componentes (fuente, filtros ópticos, acopladores, divisores de haz) y sus tolerancias puede provocar efectos tales como derivas y sensibilidad dependiente de la longitud de onda, corregibles en parte siguiendo un proceso adecuado de calibración.

Otra circunstancia a tener en cuenta es el tiempo de adquisición. Aunque con un espectrómetro diseñado para aplicaciones que requieran tomas de medidas rápidas se puede obtener un espectro de emisión completo en pocos milisegundos, el espectrofluorímetro de laboratorio utilizado para obtener los datos utilizados en este trabajo realiza una medida mucho más lenta, excitando y midiendo en las diferentes longitudes de onda secuencialmente. Con el fin de reducir el ruido gaussiano, sobre las medidas se realizaron integraciones largas. Un posible

efecto colateral es la probabilidad de que durante todo este tiempo hayan ocurrido pequeñas variaciones a nivel celular en el cultivo bajo medida y perturbe de alguna manera la medida.

Como consecuencia de estos efectos, los espectros de fluorescencia se encuentran visiblemente afectados por el ruido, manifestándose en forma de rizado superpuesto a la curva. Su efecto relativo será mayor en las curvas con fluorescencia más baja debido a una peor relación señal a ruido.

Queda de manifiesto la necesidad de aplicar técnicas de suavizado o *smoothing*. Si la curva medida por el espectrofluorímetro en cada longitud de onda viene dada por la suma entre el valor verdadero de fluorescencia y un valor de ruido, el *smoothing* trata de estimar los valores verdaderos a partir del valor medido a esa frecuencia y en bandas cercanas. Al hacerlo se ha de cuidar de que el suavizado no sea excesivo, eliminando información sutil que pudiera ser útil para la discriminación. Al fin y al cabo lo que estos métodos realizan es un filtrado paso bajo de los datos, atenuando las componentes de más alta frecuencia, asociadas con el rizado superpuesto.

3.4.1 Media móvil

Las técnicas más extensamente usadas son el método de la Media móvil (Moving average) y el de Savitzky-Golay. El primero consiste en calcular cada nuevo valor de la curva como la media de los valores contenidos en una ventana, con valores uniformes en el caso más simple, que se va desplazando a lo largo de la misma hasta haberla barrido por completo. Otra forma de entenderlo es como una convolución entre la señal y la ventana utilizada, o como el producto entre sus transformadas de fourier. En el caso de que ésta sea de forma rectangular, su transformada es la señal *sinc*, la cual tiene un decaimiento de sus lóbulos relativamente lento además de un comportamiento oscilatorio indeseado.

El parámetro ajustable para este suavizado es el tamaño de la ventana, es decir, el número de puntos de la curva que contribuirán en cada estimación. Dada la dualidad tiempo-frecuencia, una ventana grande significa un mayor filtrado paso bajo y el riesgo de excedernos, mientras que una ventana que da cabida a escasos puntos suavizará levemente. En algunas aplicaciones, como en el suavizado de curvas pertenecientes a series temporales, la ventana se escoge descentrada, o sea que la estimación de cada punto depende de su valor y un cierto número de muestras pasadas. En el caso de curvas espectrales, esta aproximación carece de sentido puesto que la influencia que las bandas vecinas tienen sobre el valor medido a una longitud de onda, es de esperar que sea simétrico, por lo que solo se aplicarán ventanas centradas.

3.4.2 Savitzky-Golay

Savitzky-Golay (Savitzky & Golay 1964) sustituye cada valor medido de la curva por el que se obtiene al hacer una regresión polinomial local con otros puntos de su entorno realizando el cálculo de los coeficientes mediante el método de mínimos cuadrados. Los parámetros a ajustar en este caso son el número de puntos a los que se tiene que ajustar el polinomio así como el orden del mismo, con la restricción de que este último sea menor al primero. También se puede utilizar para estimar las sucesivas derivadas de la curva (apartado 4.2.2).

3.4.3 Kernel

Si en lugar de una ventana uniforme usamos una función simétrica que de menos peso relativo a los puntos que se encuentran más alejados, estaremos haciendo uso de un Kernel smoother. Dentro de esta definición tan general, Moving average se puede considerar como un caso particular de esta técnica. *Kernel* es el nombre que reciben este tipo de funciones y entre sus aplicaciones se encuentran la estimación de densidades de probabilidad no paramétricas y el suavizado de curvas.

Uno de los ajustes de este método es la función a utilizar. Típicamente se utiliza el kernel gaussiano, aunque computacionalmente problemático debido a que nunca se anula por completo. El tamaño de la ventana y la desviación típica de la curva fijarían la intensidad del suavizado.

3.4.4 Wavelet denoising

El último método que se va a comparar es la reducción de ruido (*denoising*) mediante la transformada wavelet discreta (Taswell 2000). Esta transformada consiste en analizar la señal efectuando sucesivos filtrados que se centran en diferentes bandas de frecuencia. Se comienza generando dos nuevas señales a partir de la original, una con el contenido de bajas frecuencias y otra con las altas. La primera es la llamada aproximación de la señal, mientras que las componentes de la segunda reciben el nombre de detalles. En cada etapa intermedia se extraen nuevamente las altas frecuencias de la correspondiente aproximación y se van acumulando para conformar el vector de coeficientes. Dada la presencia de este instrumento de procesado en el PFC, se ofrece una explicación un poco más detallada en el tema de transformaciones, incluyendo algunos aspectos importantes que se han pasado por alto.

Los coeficientes de cada nivel nos dan una visión desglosada de la señal, lo que permite conocer donde se concentra la mayor parte de la energía de la misma, tanto en el tiempo como en

la frecuencia. Su principal propiedad es que para los primeros niveles se cuenta con un mayor número de coeficientes que proporcionan una mayor resolución temporal (que en nuestro caso se correspondería con el eje de longitudes de onda), aunque como contrapartida la resolución en frecuencia es baja debido a que sostienen un mayor número de frecuencias que los niveles superiores. De forma dual, estos últimos representan una banda de frecuencias más baja pero su reducido número de muestras impide la localización exacta donde se manifiestan.

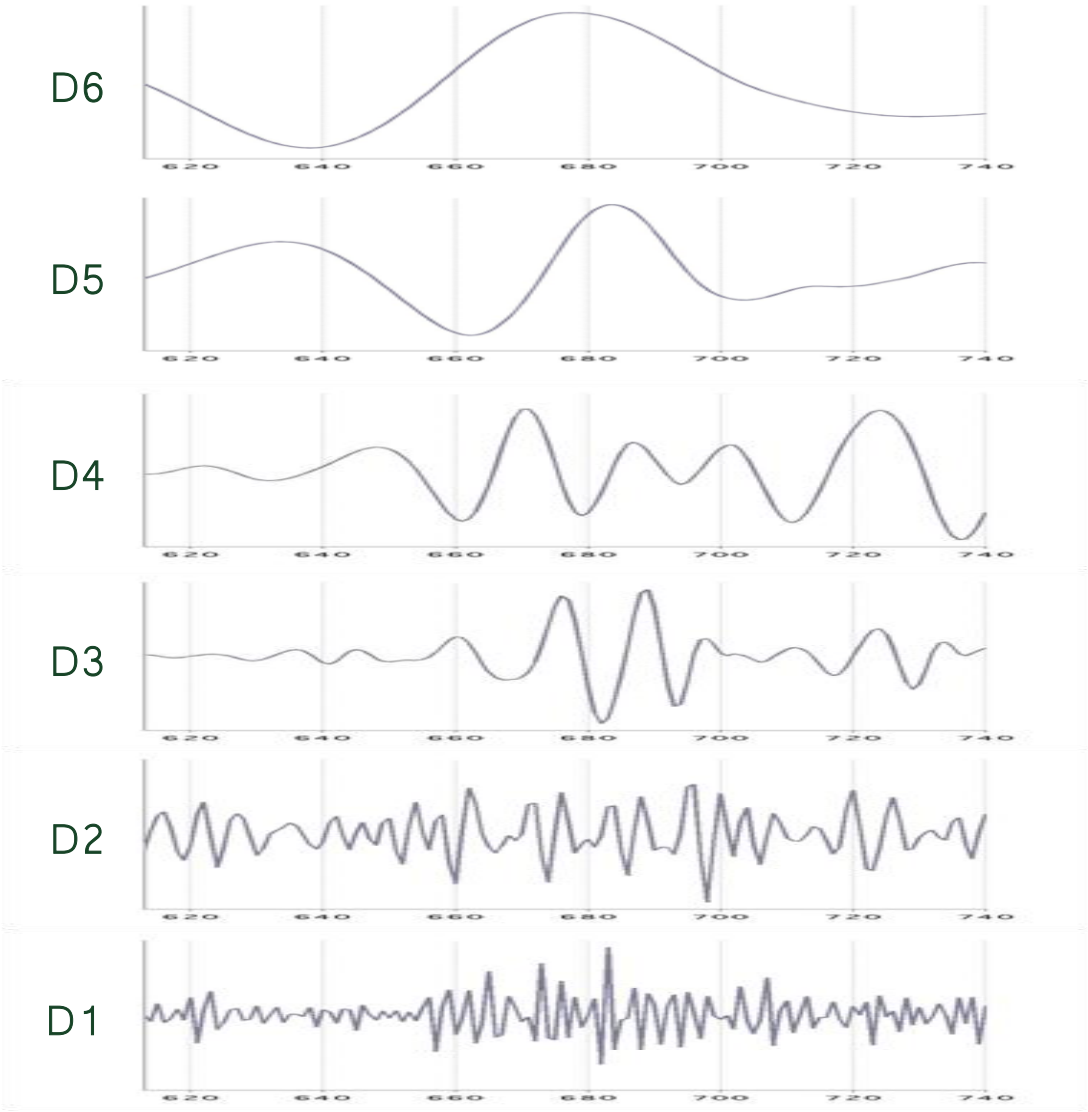


Figura 3.5. Coeficientes de descomposición wavelet de una muestra de Thwi con seis niveles de detalles con un wavelet de daubechies 9.

Para la descomposición wavelet se aumentaron las 101 bandas de trabajo a 128, que al ser potencia de dos, facilitará las sucesivas particiones por la mitad de la señal. Con este número de

muestras se puede llegar a una descomposición con siete niveles de detalles. Sin embargo se consideró suficiente con llegar al sexto nivel. Para el ejemplo de la figura 3.5 y para el resto de ocasiones en las que se utiliza la transformada Wavelet se empleó una función wavelet ortogonal de la familia de funciones de Daubechies, en concreto la D18 con un número de muestras igual al que indica su nombre.

Como se aprecia, el ruido se ve reflejado por coeficientes de pequeña magnitud especialmente visible en los primeros niveles, donde su presencia es más notable respecto a la señal. Esta propiedad es la que hace posible la reducción de ruido, eliminando aquellos coeficientes que se encuentren por debajo de un umbral escogido y que por tanto no se consideren relevantes para la formación de la parte real o deseada de la curva. El procedimiento general es el que sigue:

- Efectuar la transformada wavelet discreta de la señal.
- Calcular o estimar el umbral adecuado.
- Eliminación y ajuste de coeficientes
- Transformada inversa de los coeficientes modificados

El primer y último paso ya han sido detallados. Hay diversas formas y propuestas para calcular el umbral de forma generalizada, muchos de ellos basados en la estimación de la desviación estándar del ruido. La principal distinción es entre las que utilizan el mismo umbral para todos los niveles, como VisuShrink (Donoho 1995), o las que estiman uno diferente para cada descomposición, como por ejemplo SureShrink (Donoho & Johnstone 1995).

Una vez estimado un umbral, el nuevo valor de los coeficientes que queden por debajo de él será cero, mientras que para los que quedan por encima hay dos opciones. O bien conservan su valor o son reducidos, pasando a ser el eje que define el umbral la nueva referencia de coeficientes nulos. Éstos son conocidos como umbral duro y suave, respectivamente.

Las pruebas realizadas con algunos estimadores no dan muy buenos resultados. Al ser generalistas para poder ser utilizados en diversas aplicaciones, en nuestro caso han resultado ser o muy conservadores o muy agresivos a la hora de quitar coeficientes. El problema es que, o bien no actúan lo suficiente sobre los primeros coeficientes de detalles, o quitan información útil de los últimos.

Por ello se optó por hacer una selección del umbral ajustada al problema analizando los datos disponibles con ayuda de la interfaz gráfica de la wavelet toolbox de matlab. Como ya vimos, el ruido es más evidente en los primeros niveles de la descomposición. A partir del tercer nivel comienza a vislumbrarse un patrón que en mayor o en menor medida es compartido por las muestras. Se decidió eliminar por completo los coeficientes de los dos primeros niveles, *d1* y *d2*, y fijar un umbral suave en los restantes. El umbral de estos últimos se calcula estimando la desviación estándar de ruido a partir de la mediana de los coeficientes.

3.4.5 Selección de modelo y parámetros

Con esto hemos logrado parte de nuestro objetivo, eliminar el ruido de las muestras espectrales. Sin embargo falta cuantificar de algún modo los daños que se están produciendo sobre la forma original del espectro. No es fácil encontrar un criterio objetivo para decidir el método a utilizar y cómo ajustar los parámetros. Vaiphasa (2006) expone que los filtros de suavizado pueden causar severos cambios en las propiedades estadísticas de los espectros, afectando al resultado de aquellos métodos basados en este tipo de características. En lugar de utilizar criterios subjetivos aplicados a cada problema en particular, propone un método para estudiar qué técnicas causan menor distorsión sobre las propiedades estadísticas de los datos.

Para ello se vale de un test estadístico, la prueba t pareada. A diferencia de la prueba t, que compara dos grupos de muestras independientes, la prueba pareada la realiza entre grupos cuyas muestras están relacionadas. El test realiza un contraste de hipótesis sobre si un nuevo grupo, formado por la diferencia de los dos anteriores, tiene media cero. De esta forma, la prueba es transparente al hecho de que ambos grupos tengan distinta varianza, aunque supone que éstos poseen una función de densidad de probabilidad normal.

En el caso que nos ocupa, un grupo de muestras consistiría en los valores que toma la curva espectral de una especie en particular en una longitud de onda concreta, realizándose un test por cada una de las que forman parte de la banda de trabajo. Si los datos están normalizados (necesidad que será justificada en el siguiente apartado) es de esperar que las intensidades de fluorescencia en una longitud de onda se agrupen en torno a un valor, si hay suerte, diferente para cada especie.

Cuando se suavizan las curvas, el valor en torno al cual se agrupan las muestras en cada longitud de onda no debería variar demasiado. No así la varianza, la cual se espera que sea menor al eliminar parte de las componentes ruidosas de la señal. Es por ello que el test t pareado es

adecuado, al no hacer ninguna suposición a este respecto. En cuanto a la condición de que la distribución de las variables sea normal, se realizó el test de normalidad de Kolmogorov-Smirnov (figura 3.6) para cada longitud de onda en las distintas especies para verificar qué λs verifican esta condición. La única especie que la cumple por completo es Amin mientras que el resto lo hace en mayor o menor medida. En la mayoría de los casos negativos el test no se supera porque la kurtosis de la distribución es mayor que la de la distribución normal, y no porque se haya perdido la simetría.

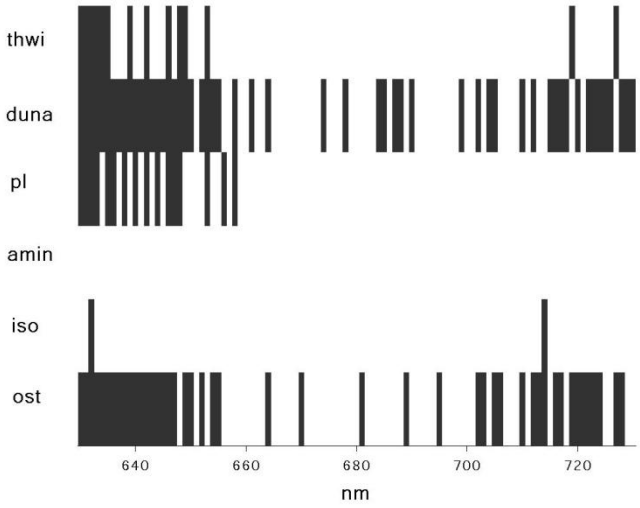
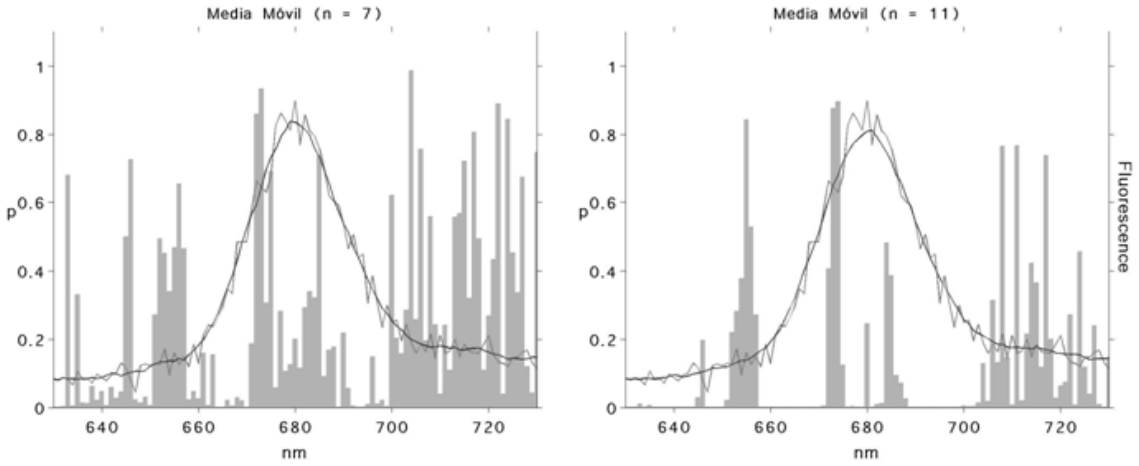


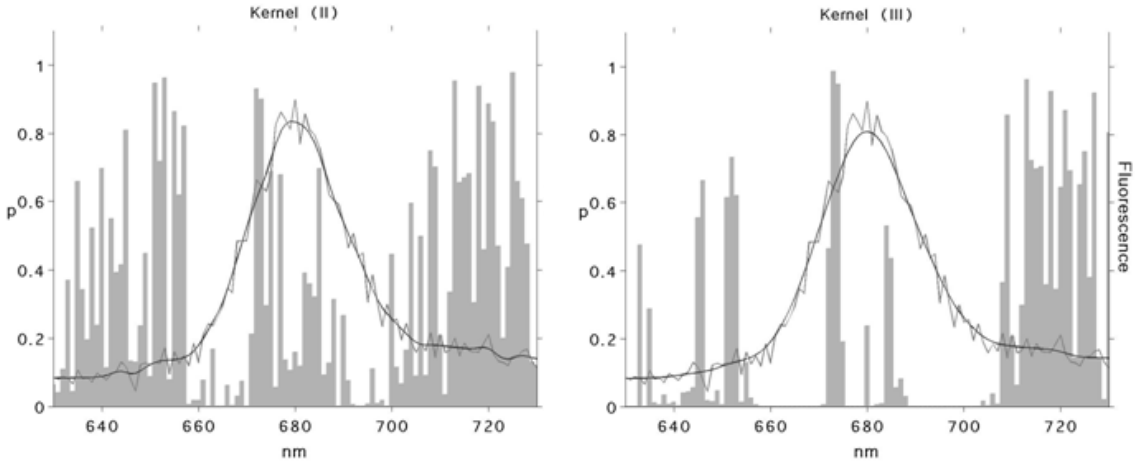
Figura 3.6. Resultado de aplicar el test de normalidad de Kolmogorov-Smirnov a las distintas especies en la banda 630-730 nm. En gris se muestran aquellas longitudes de onda para las que se ha rechazado la hipótesis de normalidad con un nivel de significación de 0.05.

Dado que se daban las condiciones, se realizó un estudio sobre cuál es el método que mejor se ajusta a la naturaleza de los datos espectrales disponibles efectuando el test t pareado entre todos los espectros de emisión utilizables y los que se obtienen como resultado de aplicar sobre ellos las técnicas de suavizado citadas con diferentes configuraciones de los parámetros ajustables. Como las muestras de la especie *Alexandrium minutum* son las que mejor han superado el test de normalidad, los resultados de ésta son más fiables y por tanto fue la que se utilizó.



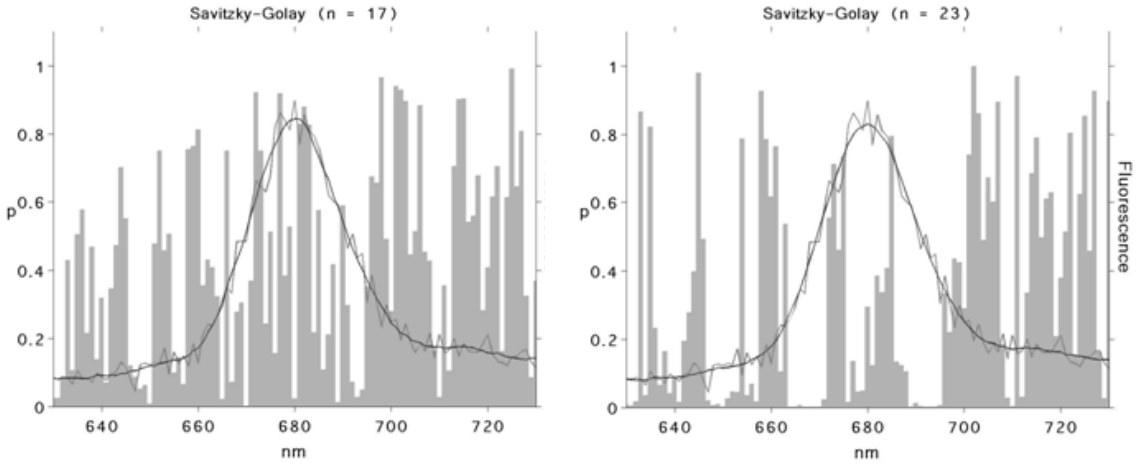
(a) (b)

Figura 3.7. Resultado de aplicar el test t pareado sobre las muestras de la especie Amin al ser suavizadas usando Media Móvil con un tamaño de ventana igual a 7 (a) y 11 (b). Las barras indican el valor p obtenido para cada λ, y superpuesta sobre ellas un ejemplo de espectro de la misma especie y su versión suavizada.



(a) (b)

Figura 3.8. Resultado de aplicar el test t pareado sobre las muestras de la especie Amin al ser suavizadas usando Kernel *smoothing* con la ventana gaussiana II (a) y III (b) (figura 3.18). Las barras indican el valor p obtenido para cada λ, y superpuesta sobre ellas un ejemplo de espectro de la misma especie y su versión suavizada.



(a) (b)

Figura 3.9. Resultado de aplicar el test t pareado sobre las muestras de la especie Amin al ser suavizadas usando Savitzky-Golay con un tamaño de ventana igual a 17 (a) y 23 (b). Las barras indican el valor p obtenido para cada λ, y superpuesta sobre ellas un ejemplo de espectro de la misma especie y su versión suavizada.

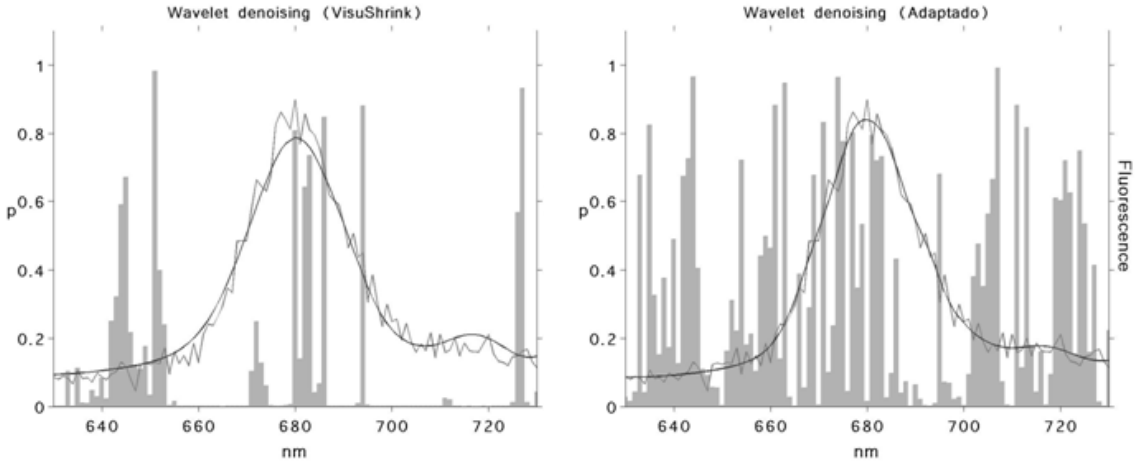


Figura 3.10. Resultado de aplicar el test t pareado sobre las muestras de la especie Amin al ser suavizadas usando Wavelet *denoising* usando estimación de umbral VisuShrink (a) y con la versión implementada (b). Las barras indican el valor p obtenido para cada λ, y superpuesta sobre ellas un ejemplo de espectro de la misma especie y su versión suavizada.

En las gráficas de las figuras de la 3.7 a la 3.10 se muestran los resultados de la prueba t pareada, teniendo en cuenta que los datos han sido previamente normalizados mediante la técnica de media y varianza que aparece en el siguiente apartado. Para cada λ aparece una barra que indica el valor de probabilidad de que la hipótesis nula (las medias de los datos antes y después de suavizar son iguales) sea cierta. La hipótesis nula es rechazada cuando el valor de p es menor 0.05

para un nivel de significación del 5%. La tabla 3.1 indica el número de bandas rechazadas por la prueba.

	Nº de bandas con p<0.05	
	(a)	(b)
Moving Average	37	66
Kernel *smoothing*	20	61
Savitzky-Golay	10	37
Wavelet *denoising*	74	26

Tabla 3.1. Resumen numérico de los resultados del test t pareado.

De todas las técnicas empleadas la que peores prestaciones ofrece es el de la Media móvil. Usando tamaños de ventana relativamente pequeños son ya muchas las bandas que no superan la prueba y bajo el criterio de ésta se consideran distorsionadas. Si se desea utilizar esta técnica sin que modifique las propiedades estadísticas de los datos el suavizado tendría que ser muy leve, conservando gran parte del ruido.

Kernel *smoothing* tiene mayor facilidad para moldear el espectro con formas suaves y curvadas pero tampoco es excesivamente tolerante al aumento de la desviación típica de la gaussiana que utiliza como ventana. Estas dos técnicas tienen una evidente dificultad para conservar la forma de zonas más pronunciadas de las curvas, en este caso el máximo y sus alrededores, quedándose visiblemente por debajo de la real. Quedan relegadas estas dos técnicas a realizar estimaciones de puntos relativamente locales con una visión limitada.

En cambio, Savitzky-Golay, permite usar un gran número de puntos para estimar el nuevo espectro, sin que esto tenga como consecuencia un cambio en la distribución de las variables. Pocas bandas son las que no superan el test cuando se utiliza una ventana de 17 puntos, lográndose además, bajo un criterio visual, un aceptable suavizado de los datos.

La inclusión de Wavelet *denoising* con estimación de umbral VisuShrink es para justificar el por qué se optó por hacer un ajuste específico para nuestros datos. No hace falta un test estadístico para darse cuenta de que la nueva curva es una versión demasiado grosera de la

original. Queda patente la prudencia con la que hay que utilizar los estimadores de umbral para los coeficientes de la transformada wavelet discreta. Cuando se utiliza la técnica adaptada descrita, los resultados mejoran notablemente, aunque un 25% de las bandas han variado su media según se desprende de los resultados del test.

Aunque la desviación que sufren las componentes espectrales de su media da idea del grado de distorsión que produce el suavizado sobre la curva, existe la posibilidad de que se haya variado la media en varios puntos de la curva estimada pero respetando la relación que en un principio había entre ellas, propiedad que es deseable conservar en los espectros procesados.

Con el fin de realizar una evaluación objetiva de este aspecto se propone hacer un estudio basado en la matriz de covarianza de las muestras. La covarianza mide la variación entre todas las componentes que forman parte de un vector de variables aleatorias. Reduciendo el concepto a un vector de dos variables, se trata de medir si éstas son dependientes, es decir, si la variación de la primera causa a su vez un cambio en la segunda. Si solo contamos con una variable la covarianza pasaría a denominarse varianza. La matriz de covarianza despliega entre sus filas y columnas la covarianza de cada variable con el resto, conformando una matriz simétrica y presentando en su diagonal principal la varianza de todas ellas.

Las longitudes de onda del espectro de fluorescencia formarían el grupo de variables aleatorias y por tanto lo que mediremos es el vínculo existente entre ellas. En el caso hipotético de que todos los espectros medidos tuvieran la misma intensidad la matriz de covarianza resultante solo tendría valores significativos en su diagonal principal, correspondiente a la varianza de ruido en cada longitud de onda. Sin embargo, de las medidas reales se obtienen curvas de diferente intensidad, que debido a la forma en las que se realizaron tiene un carácter creciente si se exponen en orden de adquisición, y variable a su vez según sea la λ de excitación de la que proviene. Es evidente que el aumento de la fluorescencia tomada a una λ se reflejará también en las otras, en mayor o menor medida dependiendo de si nos encontramos en la zona de máxima respuesta a la excitación o alejados de ella. Para que cada especie conserve la forma de su espectro para diferentes magnitudes de respuesta, la forma de variar con ella también debe ser particular de cada una. Esto es lo que queda plasmado en la matriz de covarianza de las especies (figura 3.11) y por ello para su creación se utilizan los datos sin ningún tipo de tratamiento o normalización.

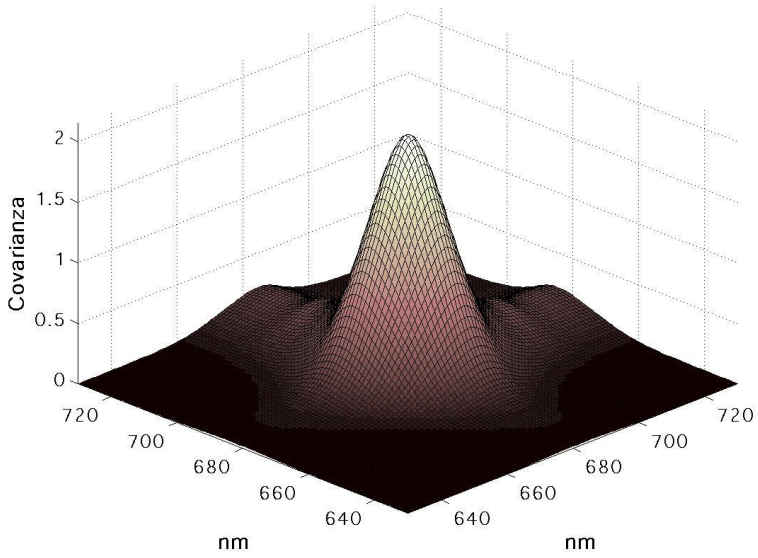


Figura 3.11. Representación en tres dimensiones de la matriz de covarianza de la especie Pleurochrysis elongata (Pl).

Las primeras diez λs no ofrecen apenas variación con la intensidad mientras que la que posee mayor varianza para esta especie se encuentra alrededor de 684 nm. Todas las filas tienen la misma forma, aunque con diferente magnitud. Por este motivo, la relación entre las variables puede quedar caracterizada por una de ellas, por ejemplo la de mayor amplitud.

La forma del espectro de emisión de las diferentes especies plasmada también en su curva de covarianza, coincidiendo a su vez los máximos de uno y otro (recordemos sin embargo que los espectros de Iso tienen una especial tendencia a desplazar su máximo coincidiendo con un aumento de la intensidad de fluorescencia). La curva de covarianza de Amin sufre cierta distorsión debido al menor número de muestras válidas que se poseen de ella dado que en general son de baja intensidad. Una vez más queda patente la aparente dificultad de distinguir entre las especies de fitoplancton a partir de su espectro de emisión de fluorescencia teniendo en cuenta la similitud y cercanía que hay entre ellas.

Una vez que disponemos de las curvas de covarianza de las muestras sin procesar tenemos una referencia sobre la cual medir si una técnica de *smoothing* está llevándose por delante parte de la información. Al aplicar una técnica de suavizado y calcular la curva de covarianza de las nuevas muestras se realizará una comparación con las originales. La exigencia será que las curvas se

solapen lo máximo posible, lo cual dará a entender que la relación general entre las variables se ha conservado.

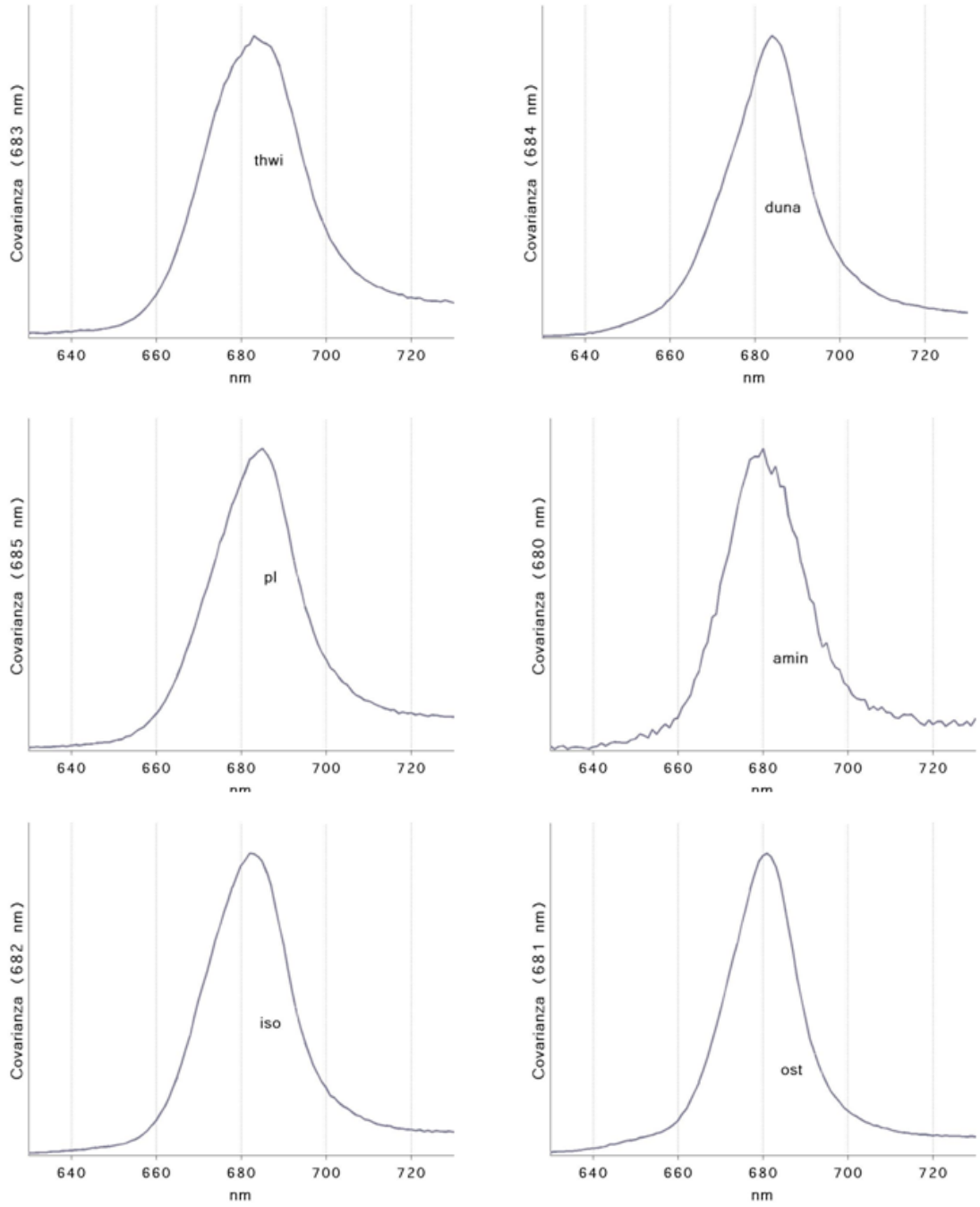


Figura 3.12. Representación de las filas para las que se obtiene el máximo de la matriz de covarianza de cada especie. La forma es un fiel reflejo de la que poseen las propias muestras individuales.

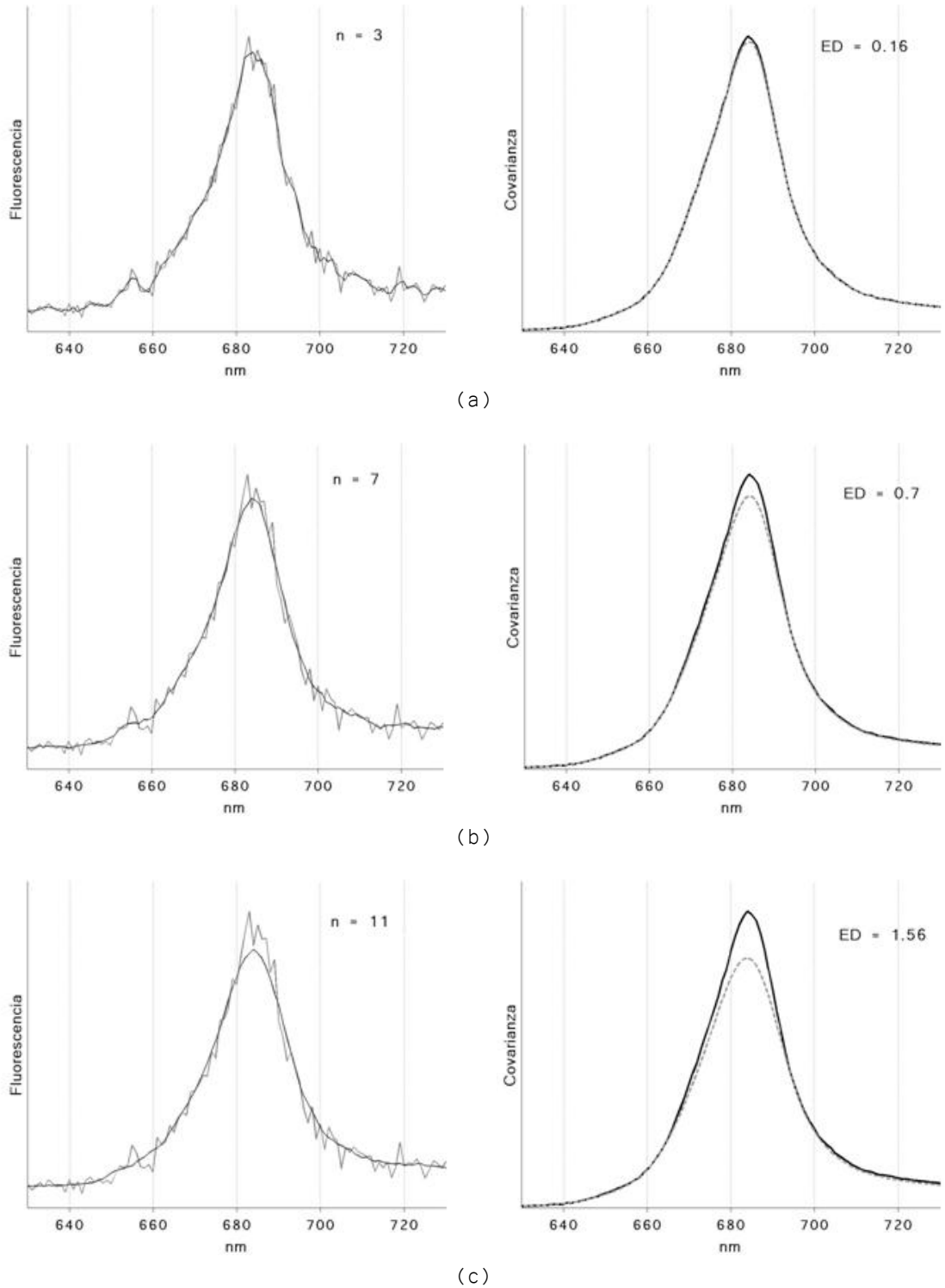
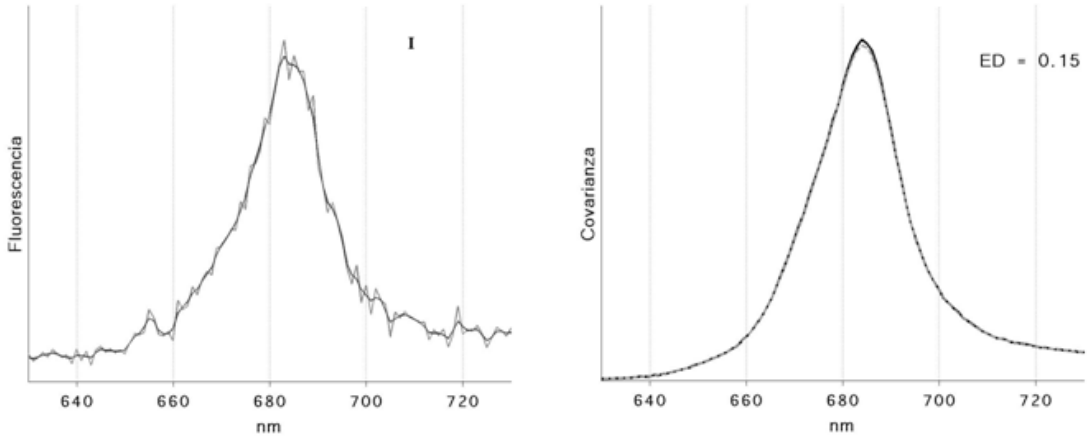
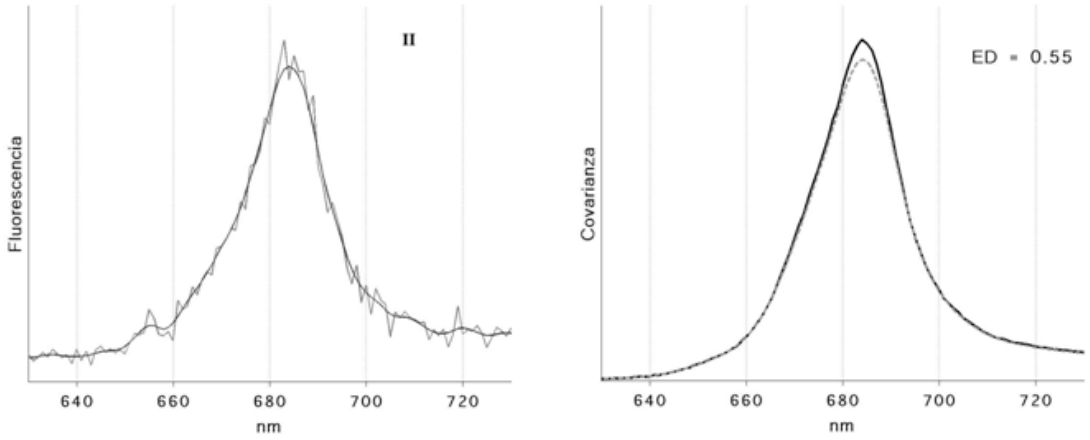


Figura 3.13. Suavizado mediante media móvil (la muestra de la especie Duna original en gris claro y la suavizada en gris oscuro) y la comparación con la curva de covarianza de las muestras sin tratar (negro continuo) y después de suavizar (gris discontinuo) para un tamaño de ventana de 3 (a), 7

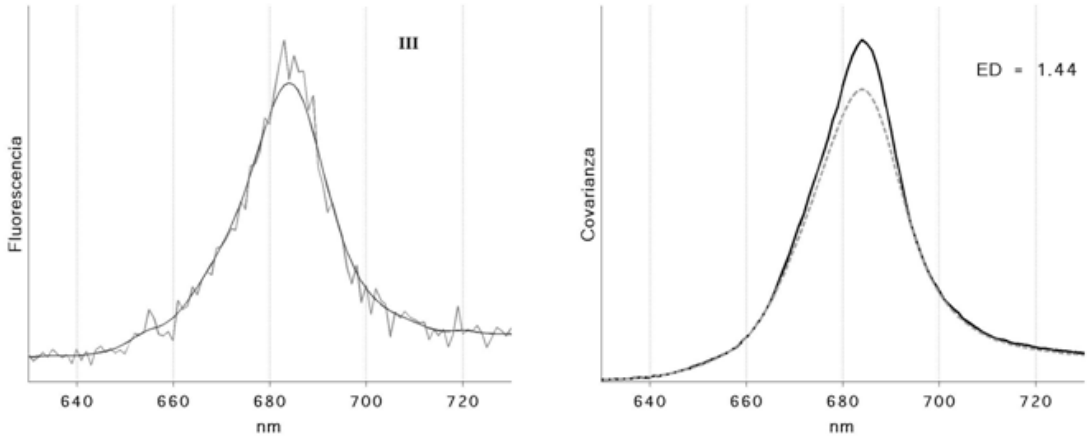
(b) y 11 (c). Junta a las curvas de covarianza figura la distancia euclidiana entre ellas.



(a)



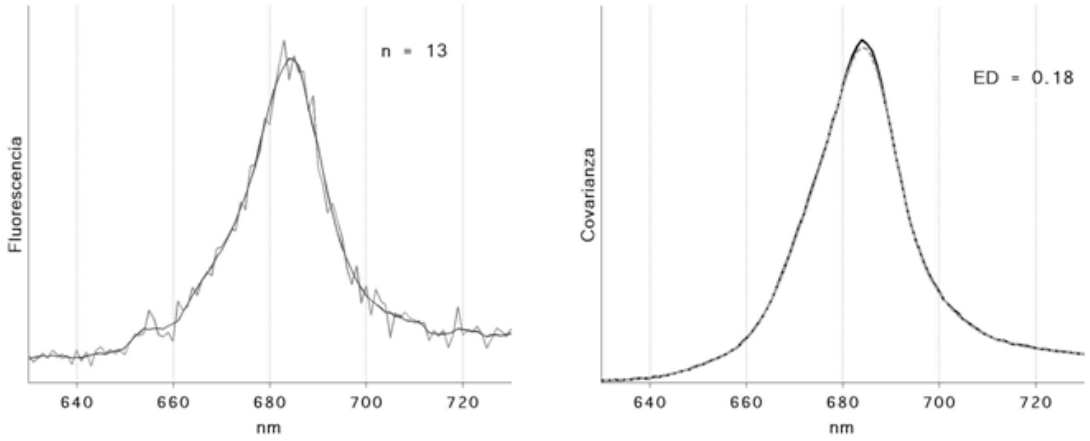
(b)



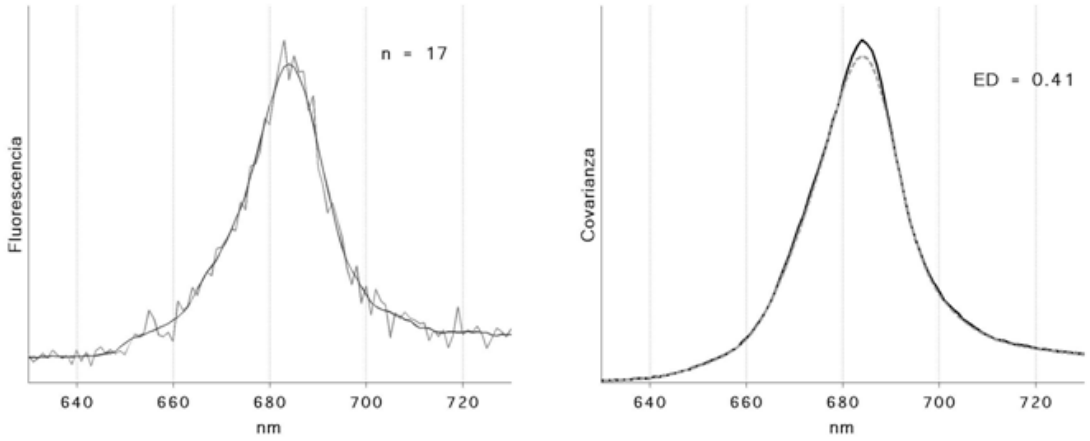
(c)

Figura 3.14. Suavizado mediante kernel gaussiano (la muestra de la especie Duna original en gris claro y la suavizada en gris oscuro) y la comparación con la curva de covarianza de

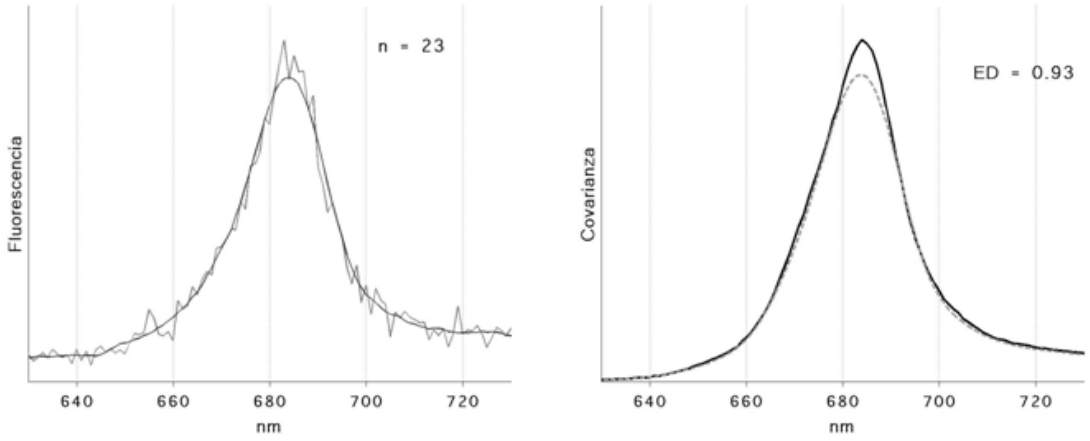
las muestras sin tratar (negro continuo) y después de suavizar (gris discontinuo) usando el kernel I (a), II (b) y III (c). Junta a las curvas de covarianza figura la distancia euclidiana entre ellas.



(a)

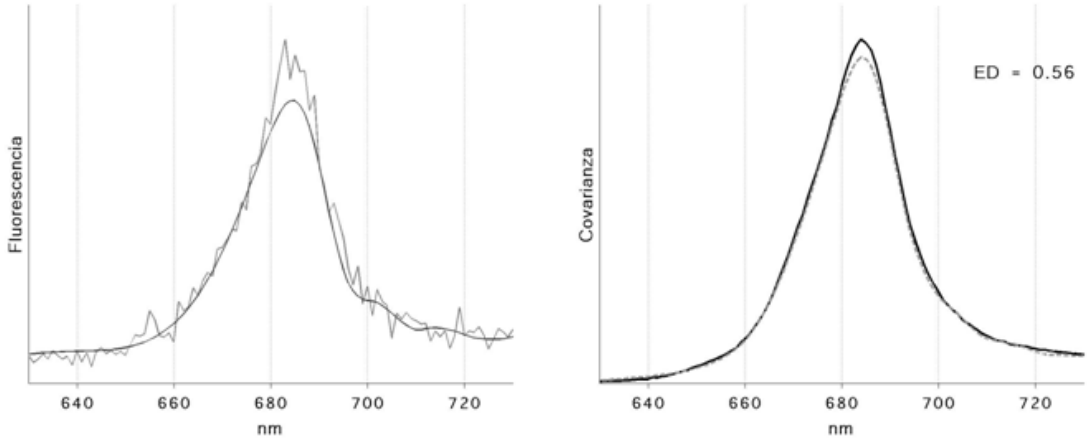


(b)

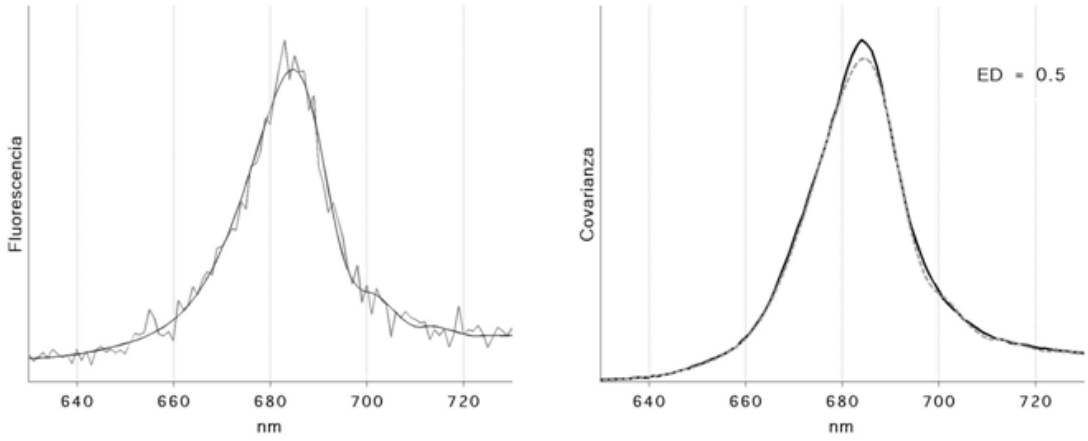


(c)

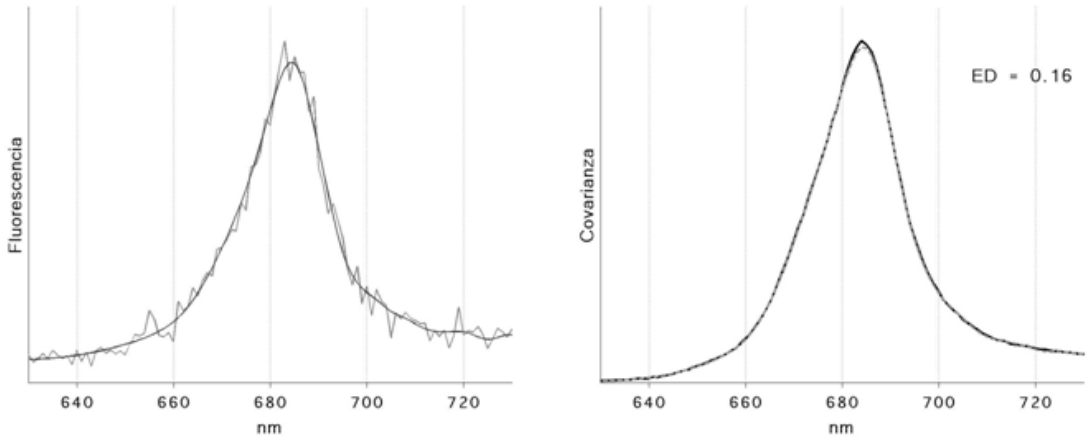
Figura 3.15. Suavizado mediante Savitzky-Golay (la muestra de la especie Duna original en gris claro y la suavizada en gris oscuro) y la comparación con la curva de covarianza de las muestras sin tratar (negro continuo) y después de suavizar (gris discontinuo) para una ventana de 13 (a), 17 (b) y 23 (c) y usando polinomios de orden 2. Junta a las curvas de covarianza figura la distancia euclidiana entre ellas.



(a)



(b)



(c)

Figura 3.16. Suavizado mediante wavelet (la muestra de la especie Duna original en gris claro y la suavizada en gris oscuro) y la comparación con la curva de covarianza de las muestras sin tratar (negro continuo) y después de suavizar (gris discontinuo) usando VisuShrink como estimador del umbral (a), eliminando los tres primeros niveles de detalles (b) y usando un método ajustado al problema (c). Junta a las curvas de covarianza figura la distancia euclidiana entre ellas.

En los ejemplos de las figuras de la 3.13 a la 3.16, se muestra por un lado la comparación entre una muestra original de la especie *Dunaniella primolecta* y la versión suavizada de la misma por cada una de las técnicas empleadas, y por otro la variación de la curva de covarianza al recalcularla después de haber usado la misma técnica de suavizado con el resto de muestras de la especie. Cada técnica de *smoothing* se probó con diferentes configuraciones de los parámetros que las definen. Se seleccionó la especie Duna para mostrar los resultados porque su forma particularmente estrecha y puntiaguda la hacen más proclive a desnudar los defectos de las técnicas.

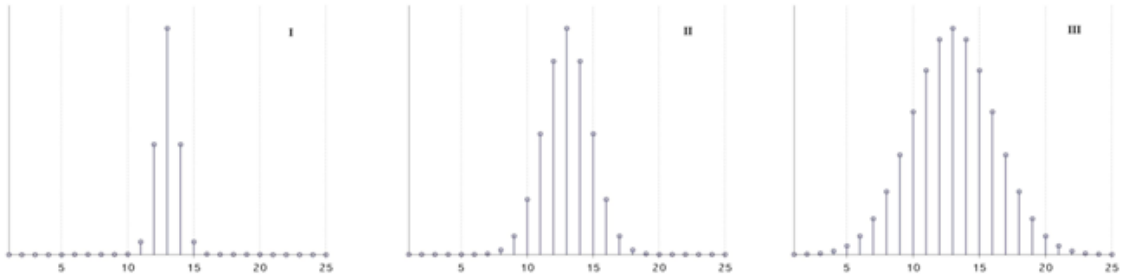


Figura 3.17. Ventanas gaussianas utilizadas para aplicar el suavizado mediante kernel.

La tónica general es que para valores pequeños o moderados de los parámetros de ajuste de las técnicas de suavizado las curvas de covarianza presentan un alto solape. A medida que aumenta el factor de suavizado éstas comienzan a separarse, como también lo atestigua el aumento de la distancia euclidiana entre ellas.

Tanto el test t pareado como el de covarianza coinciden en señalar a la Media móvil y a Kernel *smoothing* como las opciones menos recomendables para nuestro caso. Nuevamente queda reflejada la dificultad de éstas para ajustarse a la zona del máximo de fluorescencia. Si se utilizaran sus aproximaciones de las curvas para clasificar las muestras de fitoplancton, el hecho de ensanchar el espectro de una especie como Duna puede ser suficiente para que se confundan sus muestras con las de otra de con un espectro más ancho como Thwi.

Savitzky-Golay vuelve a obtener buenos resultados, pero también queda más claro su punto débil. Para ventanas con más de 17 puntos comienza a ser notable la desviación entre los máximos de las curvas de covarianza, por tanto no se utilizarán ventanas superiores.

Se ha vuelto a incluir el *denosing* con Wavelet con estimador VisuShrink. Aunque de nuevo la curva estimada difiere bastante de la original, la desviación entre las curvas de covarianza no parece tan grave en lo que al máximo se refiere. Sin embargo, son notables las ondulaciones que aparecen en los 30 últimos nanómetros. Teniendo en cuenta que en esa zona del espectro la intensidad de fluorescencia es menor, se pueden obtener desviaciones relativas demasiado grandes.

Se incluye también otra versión de Wavelet *denoising* pero simplemente eliminando los coeficientes de los tres primeros niveles. Lo que se pretende mostrar es que la mayor parte del ruido está concentrado en estos niveles, aunque también haya ciertas componentes de importancia para la correcta reconstrucción del espectro, especialmente en el tercer nivel.

3.4.6 Fusión de técnicas de suavizado

Los mejores resultados del test de covarianza es la versión adaptada de wavelet *denoising*. No solo el ajuste entre las dos covarianzas es casi perfecto sino que la inspección visual de la curva resultante revela un espectro muy entroncado con el original. La única tara es el efecto ondulatorio que se manifiesta en las últimas bandas, aunque menos acusado que en los otros dos casos.

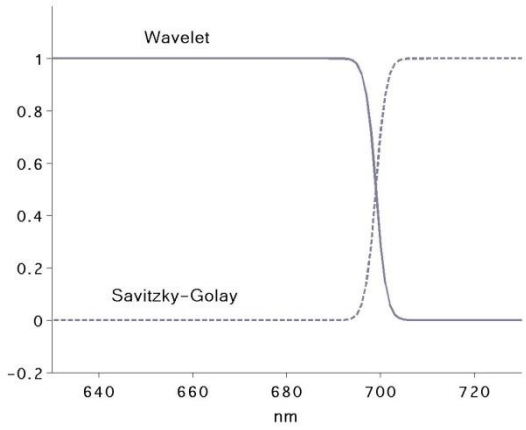
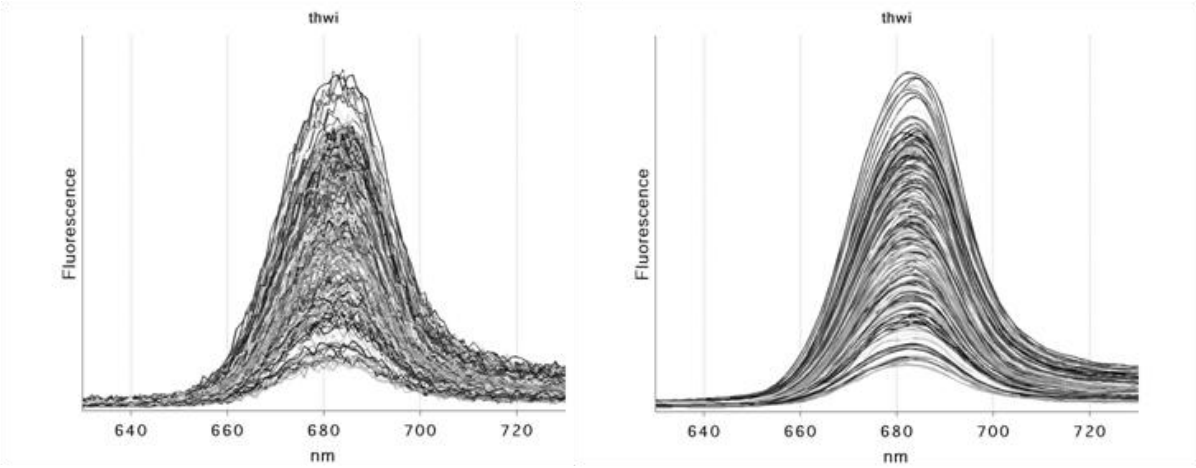
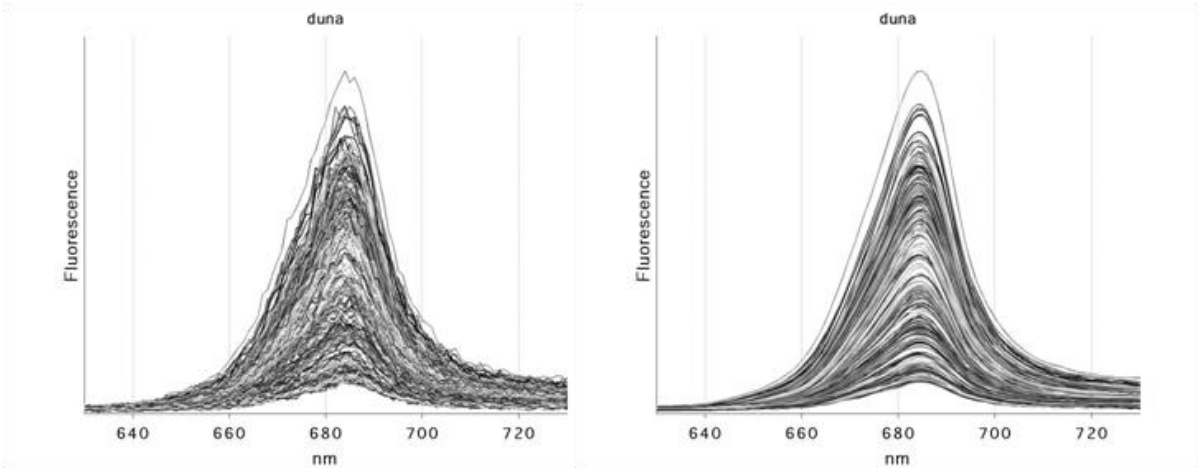


Figura 3.18. Funciones de ajuste para realizar una transición progresiva de los espectros que proporcionan Wavelet *denoising* y Savitzky-Golay.

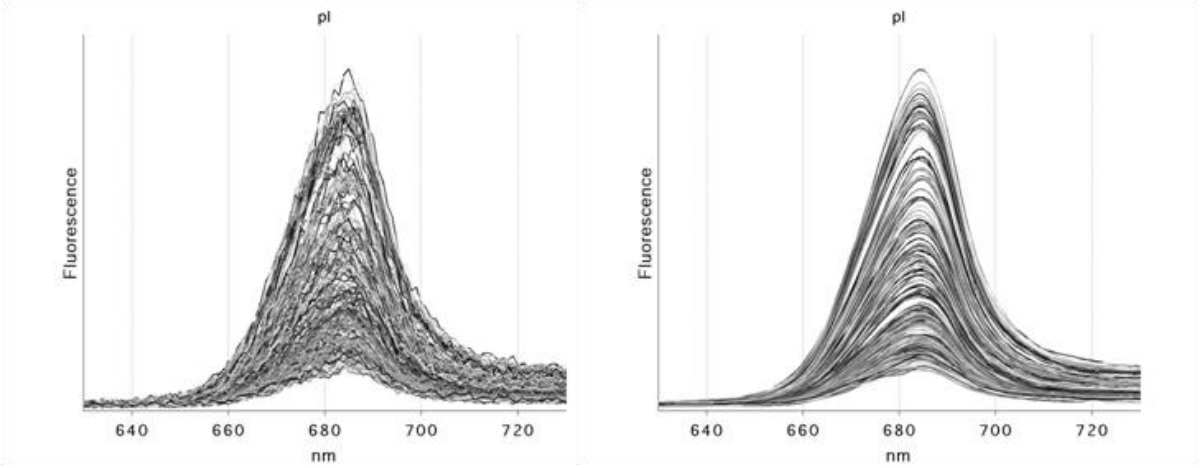
Como Wavelet denoising es el que mejor se adaptada a las zonas más pronunciadas y Savitzky-Golay también tiene un buen comportamiento y carece del problema ondulatorio se quiso estudiar la posibilidad de combinar dos técnicas de suavizado. El primero abarcaría todo el espectro excepto alrededor de las últimas 30 λs, de las que se encargaría el segundo. En la zona de fusión se cuidó de hacer un cambio de progresivo de uno a otro haciendo uso de las funciones de ajuste que se muestran en la figura 3.18. El espectro combinado se calcularía como la suma de los que proporcionan Wavelet *denoising* y Savitzky-Golay multiplicada por la función de ajuste correspondiente.



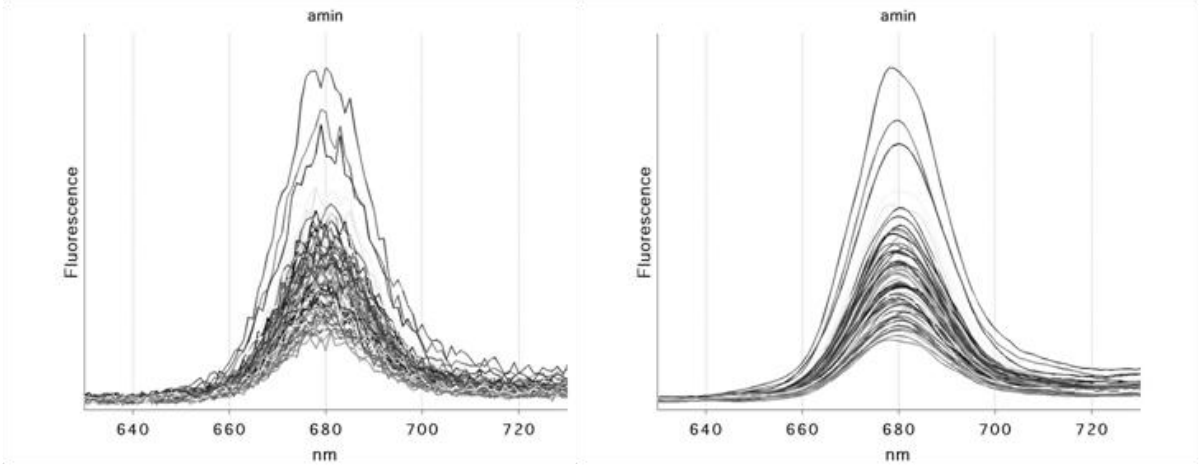
(a)


duna
Fluorescence
640
660
680
700
720
nm
duna
Fluorescence
640
660
680
700
720
nm

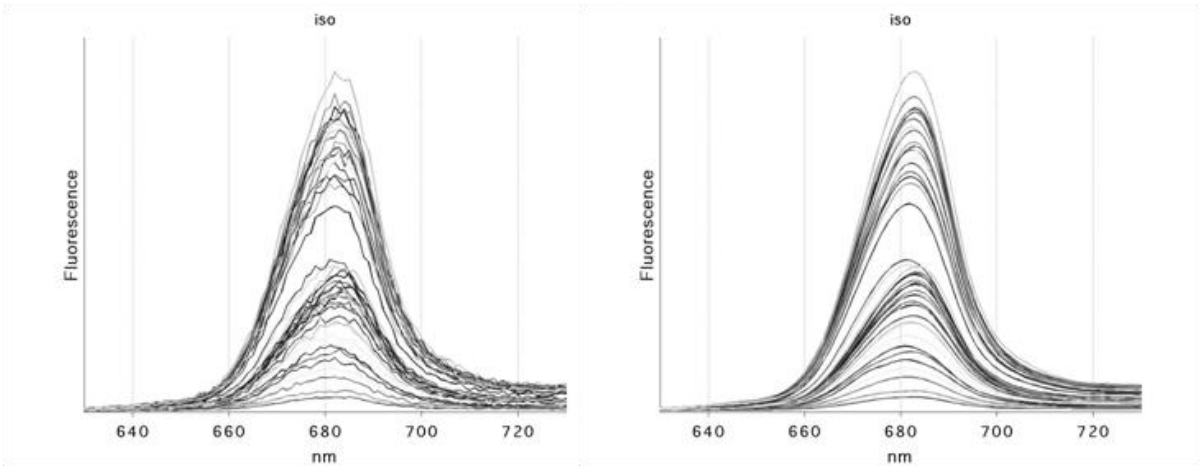
(b)


pl
Fluorescence
640
660
680
700
720
nm
pl
Fluorescence
640
660
680
700
720
nm

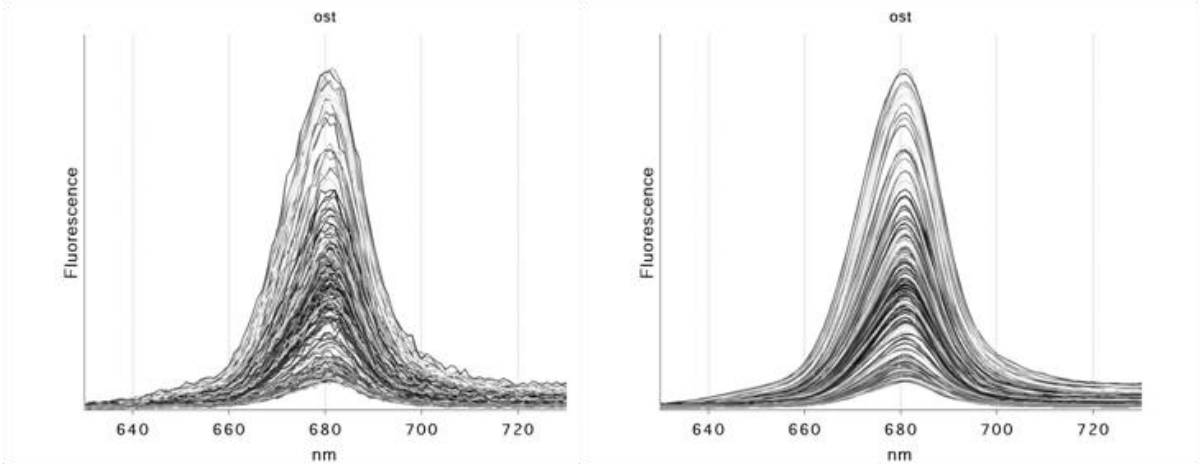
(c)


amin
Fluorescence
640
660
680
700
720
nm
amin
Fluorescence
640
660
680
700
720
nm

(d)



(e)



(f)

Figura 3.19. Resultado de aplicar la fusión de técnicas de suavizado sobre las muestras de Thwi (a), Duna (b), PI (c), Amin (d), Iso (e) y Ost (f).

3.5 Normalización de los espectros

Algunos de los factores que determinan la intensidad de fluorescencia de una muestra de agua son la concentración de células, su estado biológico o la capacidad de cada célula individual para emitir fluorescencia. Como resultado, las amplitudes de los espectros de emisión tienen un rango de variación elevado (Figura 3.20).

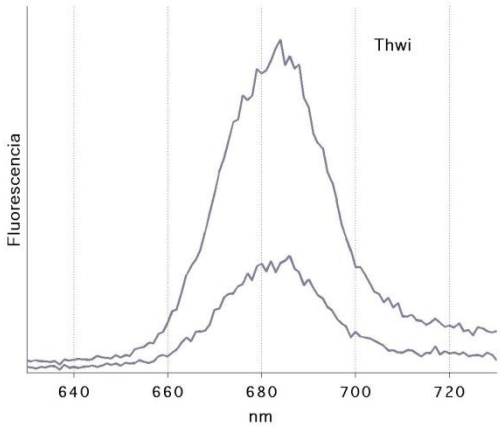


Figura 3.21. Comparativa entre el espectro perteneciente a Thwi de mayor amplitud y el de menor.

Si usáramos una medida de distancia usual, como la euclidiana, para comparar espectros de una misma especie pero de distinta amplitud el resultado sería que aparentemente existe una gran disimilitud entre ellas y sería lógico que nosotros, o nuestro clasificador supusiéramos que éstos pertenecen a especies distintas cuando en realidad no es así. De la misma forma, dos muestras de especies diferentes pero que emitieron la misma cantidad de fluorescencia, estarán relativamente próximos comparado con el resto de muestras y nuestro clasificador fracasará de nuevo en identificar la especie.

Esto pone de relieve la necesidad de igualar de alguna manera los espectros para que sea la forma que poseen y no la intensidad relativa entre ellas la que gobierne la clasificación. De alguna manera, lo que se necesita es una técnica para lograr que las muestras tengan la misma apariencia que tendrían si de todas ellas se hubiera recibido la misma intensidad de fluorescencia.

3.5.1 Máximo

La forma más evidente de lograrlo es multiplicar los espectros por un factor de escala que iguale algún parámetro de las curvas como por ejemplo la amplitud, la energía o la pendiente. Lo más habitual es fijar la amplitud máxima a la unidad, dividiendo todas las muestras de un espectro por su máximo dado que todas tienen valor positivo (Figura 3.21). Para facilitar la comparación, en este caso se han igualado todos los máximos al del espectro original de mayor amplitud.

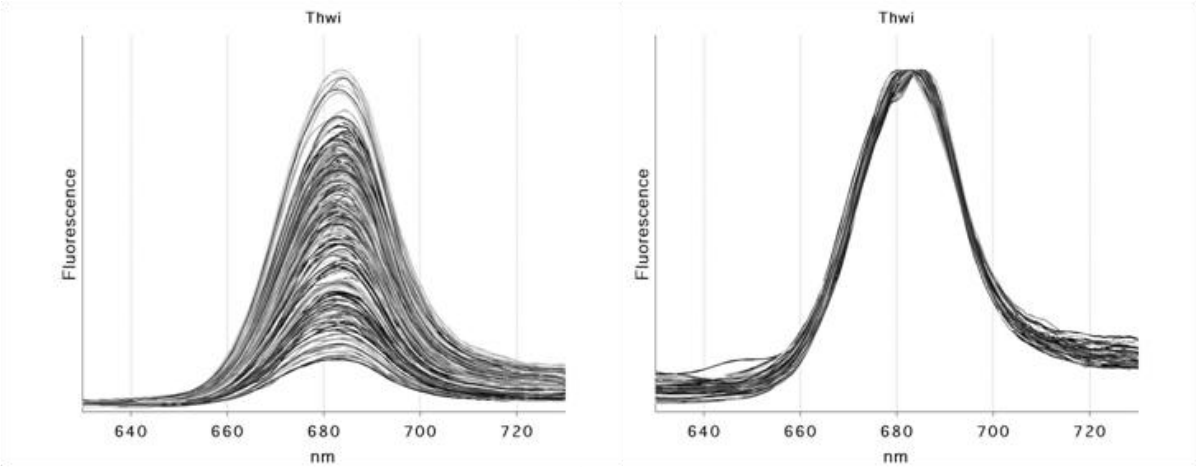


Figura 3.21. Normalización mediante escalado sobre las muestras de la especie Thwi.

El principal problema del escalado es que todas las componentes se multiplican por el mismo valor, con lo cual supone que al recibirse mayor intensidad de fluorescencia, todas las longitudes de onda aumentan de forma homogénea. Por ejemplo si buscamos que un espectro tenga la misma apariencia que tendría si hubiera recibido el doble de fluorescencia a 683 nm (su valor máximo), estimará que también habrá el doble a 660 nm o a 640 nm.

Esto se manifiesta en la figura 3.21 especialmente en las longitudes de onda con menor amplitud (las primeras 30 o las últimas 30 longitudes de onda) donde se aprecia un sobre aumento proporcional a lo pequeña que fuera su amplitud originalmente, es decir que un espectro quedará tanto más arriba cuánta más se haya tenido que escalar su máximo para llevarlo al punto de referencia.

3.5.2 Máximo y mínimo

Como las longitudes de onda que se ven más afectadas por esto son las que reciben menor fluorescencia, una posible solución es no solo llevar el máximo a un punto de referencia sino hacer lo propio con el mínimo, es decir encajar la curva en un rango fijo (figura 3.22).

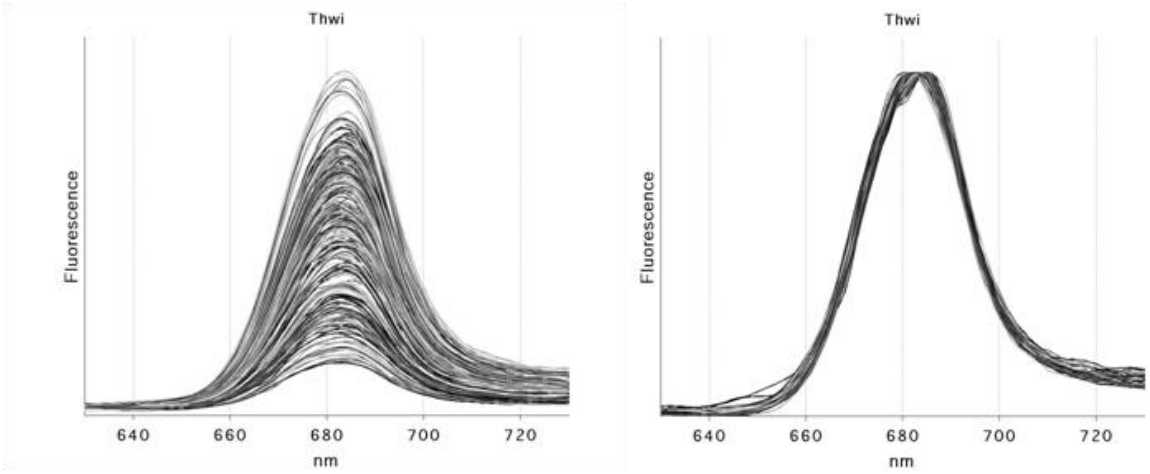


Figura 3.22. Normalización mediante doble escalado sobre las muestras de la especie Thwi.

Como se aprecia, efectivamente esta normalización logra comprimir los puntos de la curva de menor intensidad. Esto se logra porque el escalado se realiza sobre la resta de cada componente y el mínimo del espectro con lo cual cuanto más pequeño es el valor a una longitud de onda, proporcionalmente menor será el escalado. Se puede comprobar en las figuras como las componentes más intensas de la curva exhiben poca diferencia con respecto a la normalización anterior.

Sin duda esta última técnica, a pesar de su simpleza, resulta ser muy efectiva y es perfectamente utilizable para el propósito de clasificar e identificar las especies de fitoplancton. Sin embargo también presenta algún inconveniente. El efecto de escalar menos las zonas de menor amplitud de la curva, a pesar de ser acertado y deseable, no sigue ningún criterio extraído de las propias muestras por lo que no deja ser muy genérico. Por otro lado esta normalización provoca una distorsión en dos zonas concretas de las curvas: en la parte mínima y en la máxima. Todas las muestras comparten aproximadamente los puntos en los que éstos se producen, por lo que las curvas se apelotonarán en ellos sobre los mismos valores, llevándose por delante propiedades estadísticas que pudieran ser útiles para algún clasificador que hiciera uso de ellas.

3.5.3 Media y varianza

Otra técnica muy simple pero que da muy buenos resultados consiste en normalizar los espectros de forma que todos tengan el mismo valor medio y la misma varianza. El primer paso sería centrar los espectros igualando sus medias (figura 3.23).

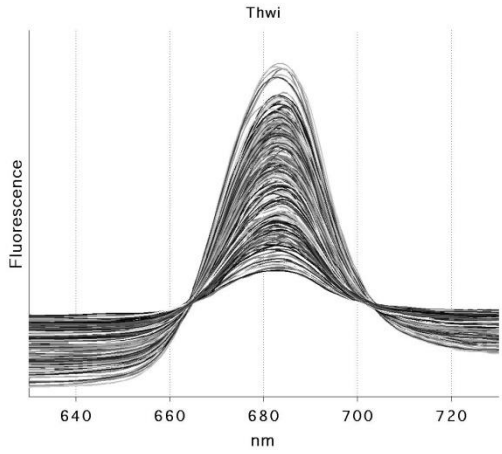


Figura 3.23. Las muestras de Thwi centradas.

La normalización se completa igualando las desviaciones estándar de las muestras. Si por ejemplo se igualan a la del espectro de mayor amplitud, las que presentan menor intensidad verán sus valores estirados, tanto más cuanto más alejados estén de la media. De esta forma se logra realzar el pico sin por ello levantar el nivel de la base (figura 3.24).

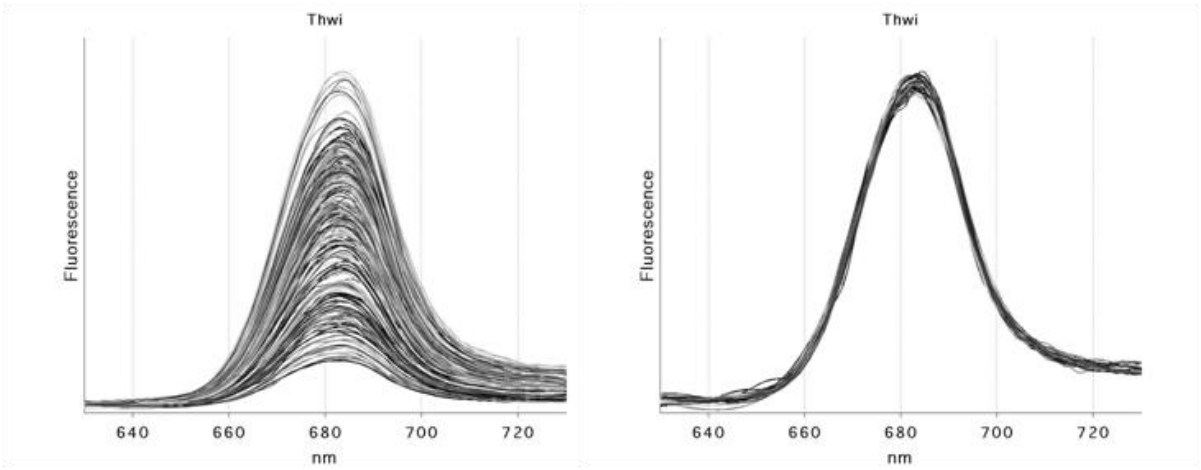


Figura 3.24. Normalización de media y varianza sobre las muestras de Thwi.

El resultado es una normalización en cierta forma más natural comparado con los otros dos casos. Es de destacar que las dos primeras normalizaciones se ven muy afectadas por el ruido que pueda haber superpuesto a la señal dado que afecta en la selección del máximo o el mínimo de la curva. Sin embargo esta última técnica es prácticamente invariante a la presencia de ruido dado

que la compensación de la desviación estándar se hace de forma equilibrada a partir de la media, cancelándose de esta manera su efecto.

3.5.4 Niveles de wavelet

Una característica común a las normalizaciones vistas es que las curvas que se desvían más del patrón general son las que provienen de señales de menor intensidad. Si nos fijamos en el último caso, al igualar los espectros al de mayor intensidad, aquellas curvas con menor relación señal a ruido son las que más tienen que variar más su forma, realzando las variaciones a pequeña escala particulares en ellas. Si procedemos de forma contraria, es decir, igualando los espectros de mayor intensidad con los de menor el resultado es el mismo puesto que en este caso estamos comprimiendo las curvas e igualmente se mantiene la proporcionalidad entre los defectos de unas y otras.

Con el propósito de utilizar un modelo que conserve las ventajas de la normalización de la media y la varianza, dígase la independencia de la normalización con el ruido y no introducir distorsión sobre la curva, pero que haga un ajuste más preciso que trate de disminuir el efecto que sobre las muestras menos intensas se produce, se implementó una técnica basada en la descomposición wavelet.

En Randolph (2006) se realizó una aproximación similar en la que se flexibilizó la normalización de la media y la varianza aprovechando las propiedades de la transformada discreta wavelet para poder realizar un ajuste más localizado. Éste se basa en la propiedad de que la varianza de la señal también sufre una descomposición al hacer la transformada para llevar a cabo la normalización a partir de la varianza de una porción de los niveles wavelet, recomponiendo la señal a partir de esta selección.

Haciendo uso de estas mismas propiedades, la normalización propuesta consiste en hacer un ajuste de la media y la varianza (para el caso de los coeficientes de la aproximación de la señal) y de la varianza (para el caso de los niveles de detalles dado que éstos tienen media nula) independiente para cada nivel. Si por ejemplo quisiéramos igualar dos espectros, tomaríamos la media y las varianzas de la descomposición de la primera y compensaríamos los coeficientes de la segunda para que tuvieran estos mismos valores. De esta forma a diferencia de lo que se propone en el citado artículo, al no fijar todas las varianzas a la unidad también se conserva la relación existente entre las de cada nivel, junto con la posible información que aporta.

Para no tener que depender del conocimiento de las propiedades estadísticas de ningún espectro en particular se propone modelar la forma en la que estas propiedades varían con la intensidad de fluorescencia y así poder modificar el aspecto de las curvas para que tengan la amplitud que se desea. Para modelar por ejemplo la dependencia de la desviación estándar del nivel de coeficientes D6 con la intensidad máxima del espectro, disponemos sus valores para todas las muestras en una representación XY. (Figura 3.25).

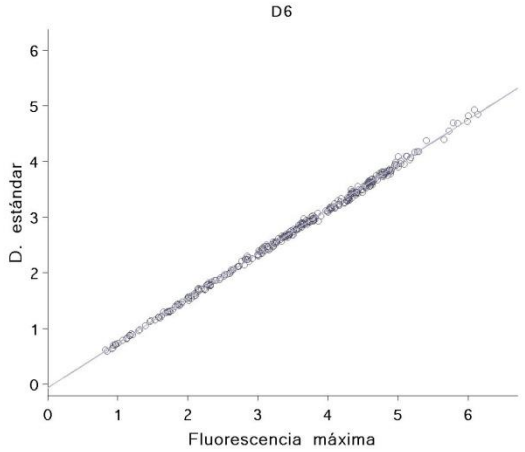


Figura 3.25. Representación de la desviación estándar de los coeficientes del sexto nivel de detalles frente a la fluorescencia máxima para la especie Thwi.

Se observa una íntima relación lineal entre las variables con poca dispersión, pudiéndose realizar una regresión lineal para modelar su comportamiento. Lo mismo ocurre con el resto de niveles con la particularidad de que a partir del nivel D3 la pendiente de la recta es prácticamente nula. En el caso de la media de los coeficientes de la aproximación A6 sucede lo mismo pero los puntos de la nube se encuentran algo más dispersos.

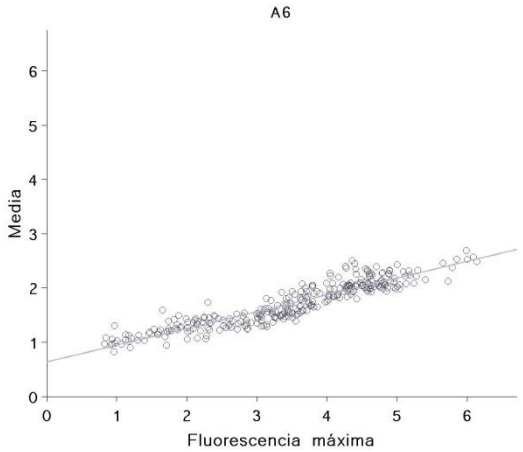


Figura 3.26. Representación de la media de los coeficientes del sexto nivel de aproximación frente a la fluorescencia máxima para la especie Thwi.

La forma de proceder de la normalización es la que sigue:

- Se selecciona el nivel de intensidad de fluorescencia máximo al que se desea normalizar las muestras.
- Se calculan los coeficientes de la transformada wavelet discreta.
- Se obtienen la media y la varianza de cada nivel de detalles y el de la aproximación.
- Para cada nivel se estiman del modelo los valores de media y varianza que tendría el espectro para el nivel de intensidad seleccionado y el que según el modelo debería tener para el nivel de intensidad actual.
- A partir de los parámetros estimados se calcula el ajuste que hay que aplicar a los que tiene la muestra.
- Se transforman los parámetros de la muestra.
- Se realiza la transformada wavelet inversa.

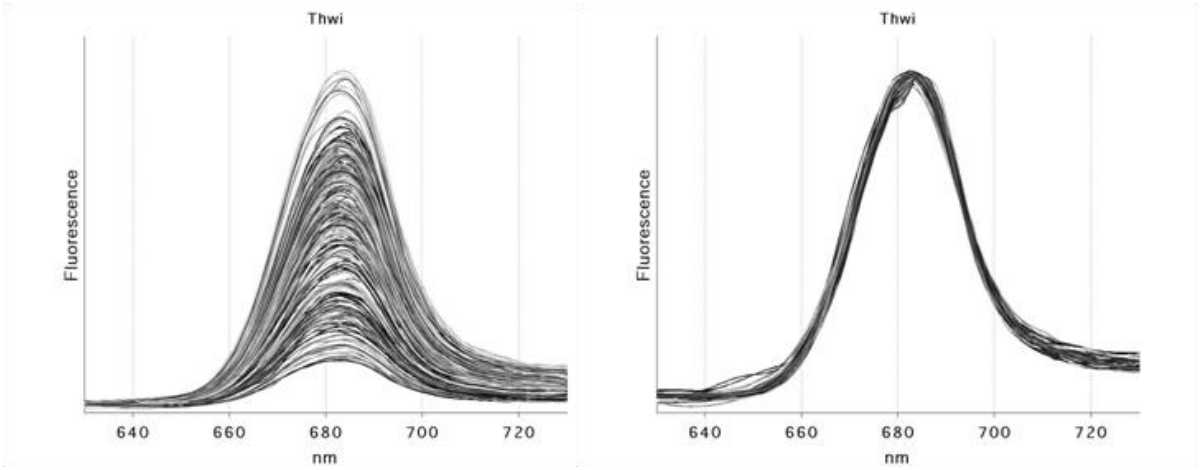


Figura 3.27. Normalización mediante la adaptación local de los niveles de descomposición wavelet.

Si comparamos el resultado de aplicar este proceso (Figura 3.27) con el que se obtuvo con la normalización de media y varianza convencional parece que las muestras estén ahora menos agrupadas aunque sí parezcan ser más uniformes. La mayor separación de las muestras viene del

hecho de que cada una adaptó sus coeficientes de aproximación con valores de media diferentes en función de la que tuviera originalmente y como vimos este parámetro es el más dispersivo del modelo aplicado.

Otra posible forma de utilizar esta normalización sin que se produzca este efecto es usar para todas las muestras la misma media para los coeficientes de aproximación, en concreto la que establece el modelo como valor típico para la intensidad de fluorescencia a la que se desea normalizar (figura 3.28).

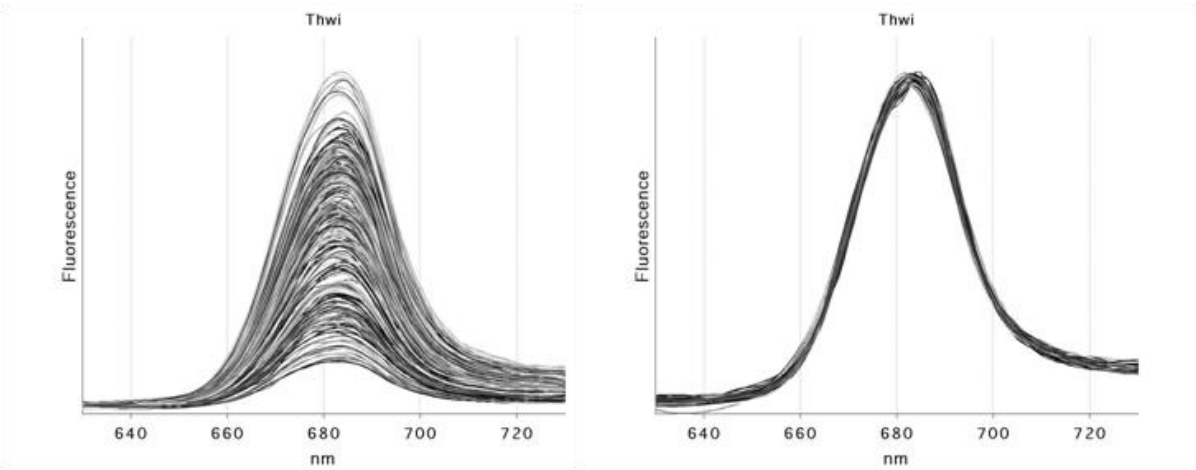
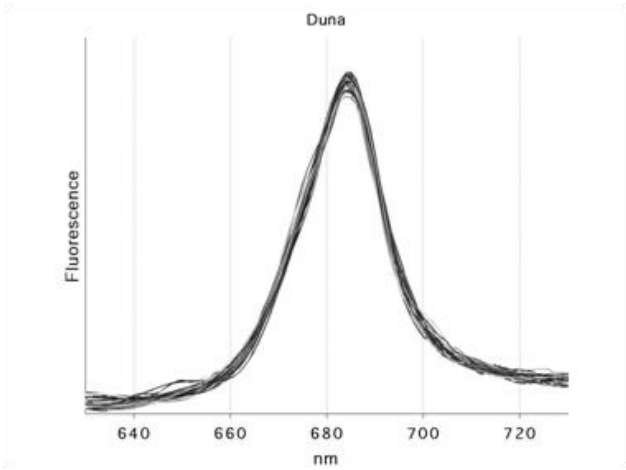
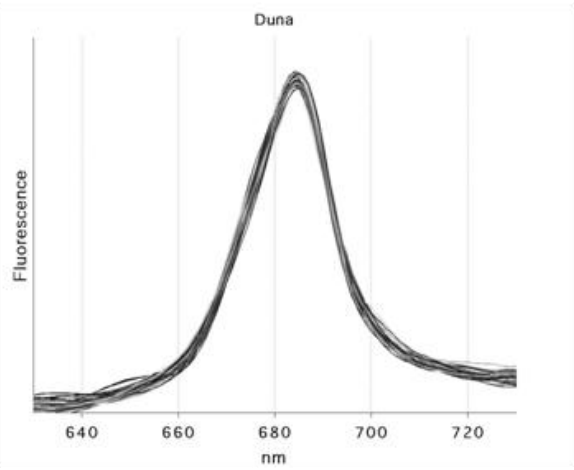


Figura 3.28. Normalización mediante la adaptación local de los niveles de descomposición wavelet.

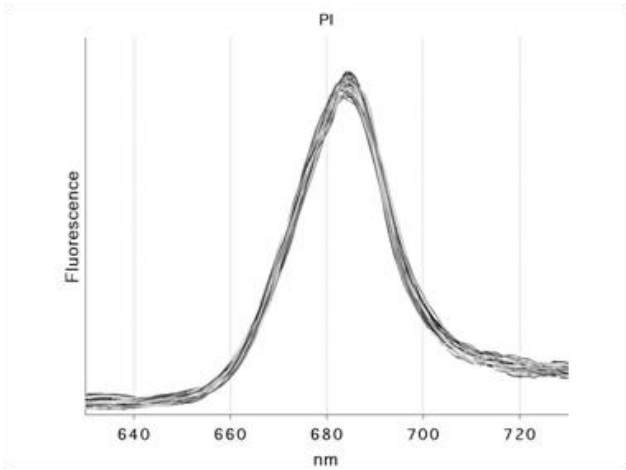
Con esta modificación se ha logrado nuevamente agrupar las muestras y además aquellas con menor intensidad presentan una menor desviación respecto al patrón general. Para facilitar la comparación entre esta última normalización y la de media y varianza se presentan también el resto de especies normalizadas siguiendo estos dos procedimientos (Figura 3.29). De nuevo se pone de manifiesto que normalizando mediante este último método se logran curvas más uniformes además de conseguir igualar mejor la zona del máximo.


Duna
Fluorescence
640
660
680
700
720
nm

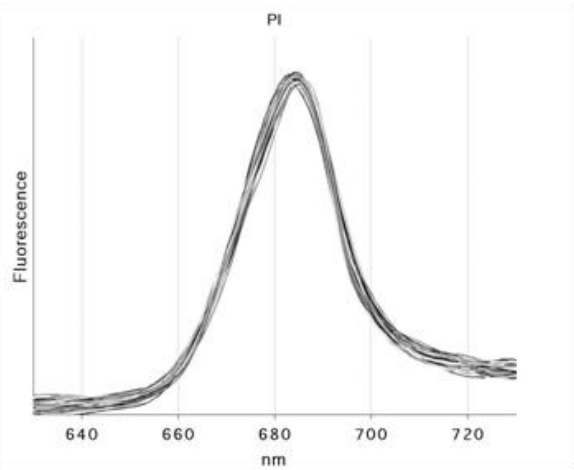
(I)


Duna
Fluorescence
640
660
680
700
720
nm

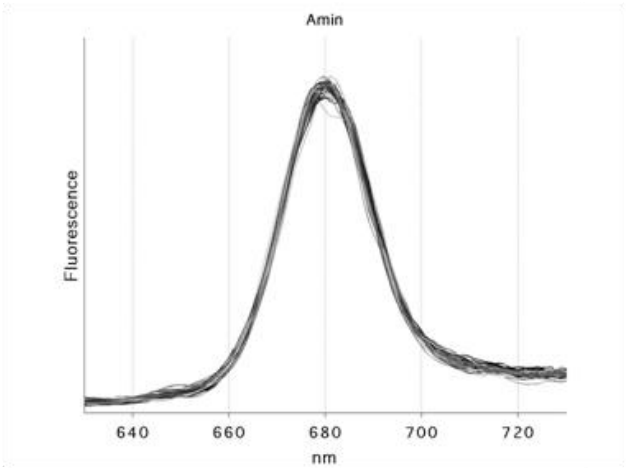
(II)


PI
Fluorescence
640
660
680
700
720
nm

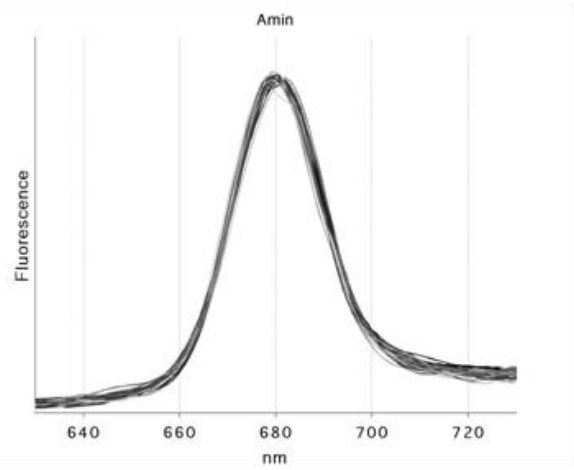
(I)


PI
Fluorescence
640
660
680
700
720
nm

(II)


Amin
Fluorescence
640
660
680
700
720
nm

(I)


Amin
Fluorescence
640
660
680
700
720
nm

(II)

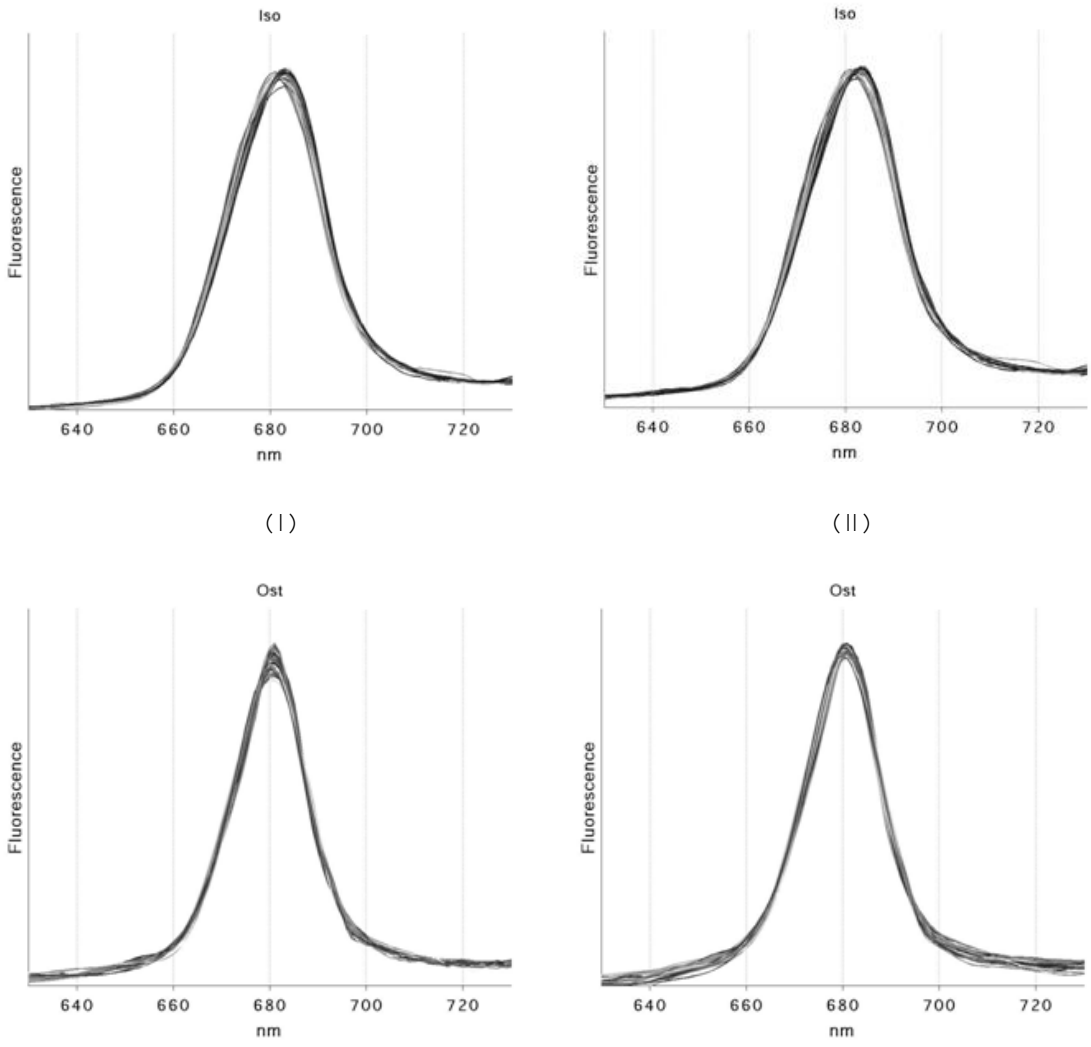


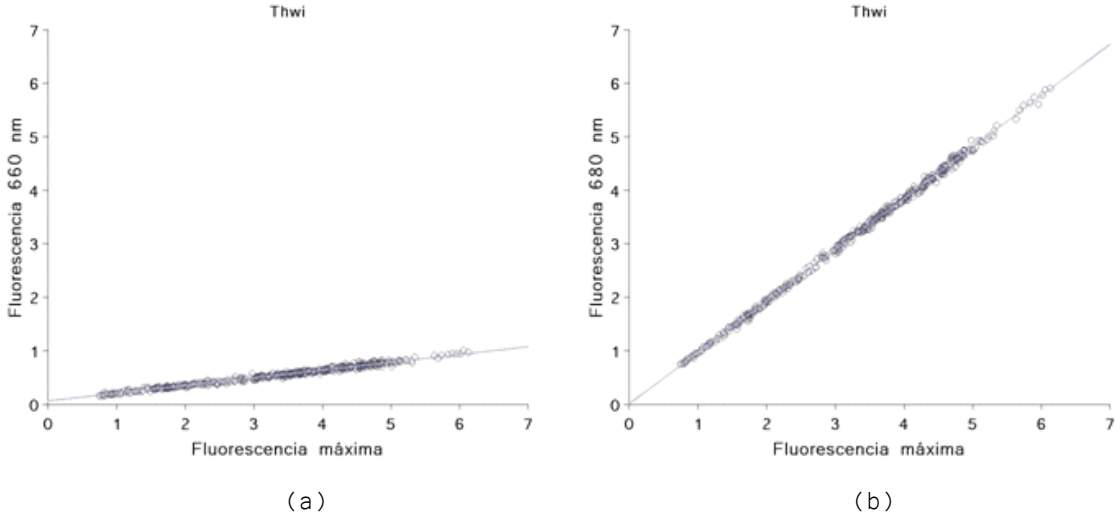
Figura 3.29. Comparativa entre la normalización de media y varianza (I) y la basada en la descomposición wavelet (II).

3.5.5 Modelado de espectros crecientes

Con esta nueva técnica se aprovechó la simpleza de la normalización de media y varianza para elaborar una que presenta mayor robustez. Se pretende ahora hacer lo propio con la normalización por escalado del máximo, desarrollando un método que trata de resolver de forma elegante sus deficiencias.

La nueva técnica se basa en el hecho de que se cuenta con un set de datos con muestras de fluorescencia que cubren un alto rango de respuestas fundamentalmente debidas a los diferentes estados en el desarrollo del cultivo bajo medida. Esta información puede ser usada para crear un

modelo de curvas crecientes para una especie determinada. Lo que se lograría es disponer de una forma objetiva de saber cuánto hay que modificar una longitud de onda concreta cuando se está adaptando el espectro completo a un nuevo valor de amplitud.



(a) (b)

Figura 3.30. Relación entre el valor máximo de fluorescencia y el que se obtiene a 660 nm (a) y 680 nm (b) para todas las muestras de la especie Thwi. Se superpone la regresión lineal calculada.

Si disponemos los valores que toman los espectros a una cierta longitud de onda frente a una de referencia como puede ser la fluorescencia máxima de cada curva, se obtiene una relación fuertemente lineal de pendiente cercana a la unidad si nos encontramos en las longitudes de onda en torno al máximo o cercana a cero si nos alejamos de él. El modelo consistirá por tanto en calcular los coeficientes de regresión lineal que se extraen para cada longitud de onda De esta forma si tenemos un espectro con una amplitud de fluorescencia de 2 unidades y queremos normalizarla para que tenga una de 5 unidades, tomaremos estos dos valores de cada recta del modelo para calcular el factor por el que debemos multiplicar el valor real medido en cada longitud de onda.

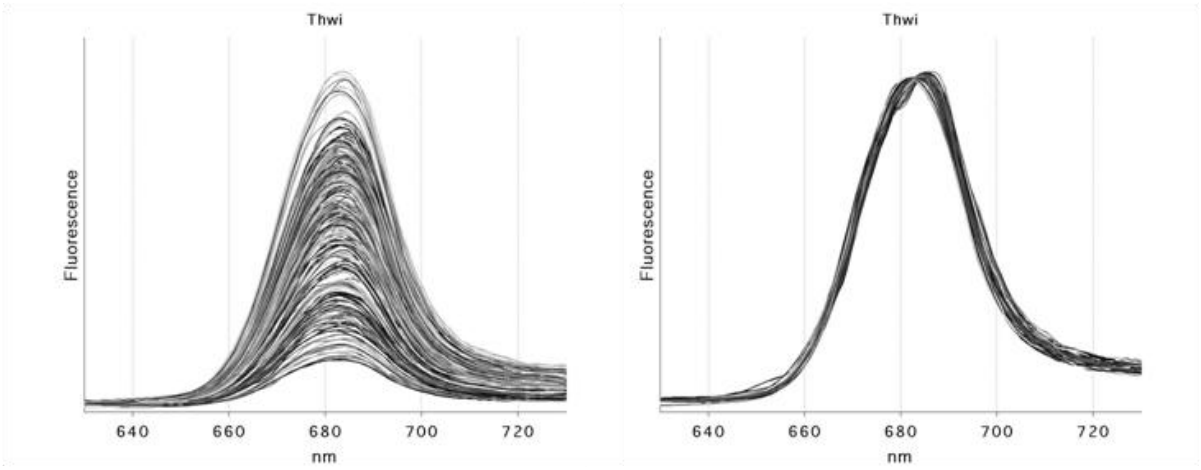


Figura 3.31. Normalización de las muestras de Thwi mediante el modelado de curvas crecientes asignando cada recta del modelo a una longitud de onda.

En una primera aproximación se realizó el modelo fijando los coeficientes de regresión a para cada longitud de onda. El problema de esta forma de proceder es que los espectros presentan cierta deriva y por tanto no están centrados. Al aplicar el modelo sobre aquellas muestras que tienen el máximo desplazado respecto de la mayoría éste aplicará a cada λ los factores de ajuste que en realidad corresponden a otra componente situada a una distancia igual al desplazamiento. En la figura 3.31 se aprecia que algunas muestras tienen el máximo por encima del resto, consecuencia de no utilizar el factor de multiplicación adecuado.

La solución es hacer un modelo relativo, por ejemplo al máximo de los espectros. Las rectas de regresión se calcularían para cada componente desplazada una cierta cantidad a la izquierda y a la derecha respecto al máximo. En realidad en la implementación efectuada se tuvo en cuenta que el valor máximo de las curvas no es más que una estimación de la posición del verdadero pico, en ocasiones no del todo preciso. Por ello se utilizaron los dos puntos que intervienen en el cálculo de la derivada (capítulo 4) para la cual ésta vale cero (máximo del espectro).

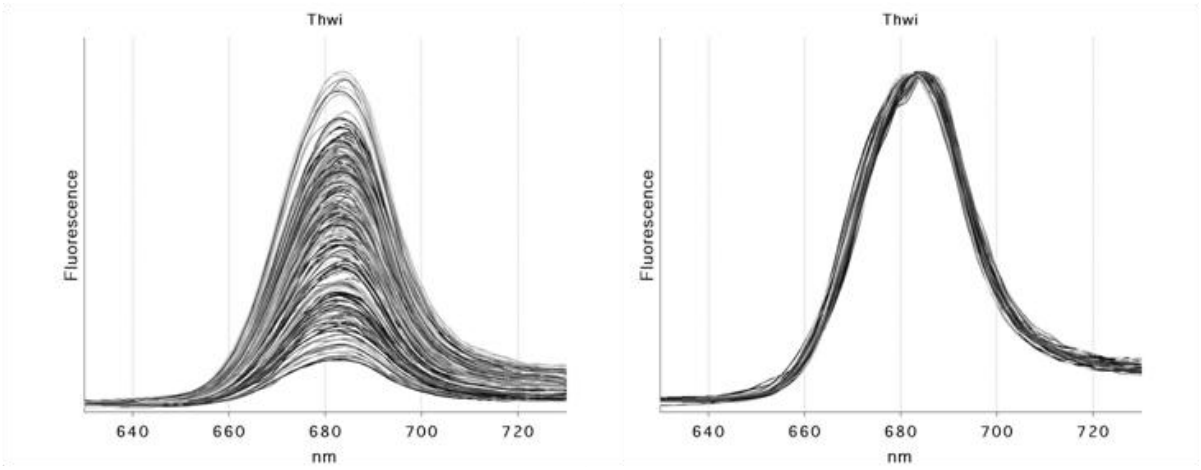


Figura 3.32. Normalización de las muestras de Thwi mediante el modelado relativo de curvas crecientes.

Comparando el resultado con el que se obtuvo con la normalización de máximo y mínimo, ésta última también distorsiona el punto máximo debido a que es este punto precisamente el que se utiliza de referencia. Sin embargo esta técnica es la que mejor normaliza la zona de bajas longitudes de onda, la cual apenas cambia con el aumento de la intensidad de fluorescencia.

Otra posibilidad que se puede contemplar es la de centrar todos los espectros mediante el mismo método empleado en la normalización relativa usando la derivada. La comparación entre especies se realizaría en base tan solo a la forma de la curva y no a otros aspectos como el hecho de que las especies tengan su máximo en diferentes λs. Sin embargo también sucede que dentro de una misma especie hay una deriva de los espectros, especialmente notable en el caso de Iso con una clara tendencia a desplazar su máximo hacia mayores longitudes de onda. Esta deriva puede perjudicar en la comparación de las muestras por lo que centrar las curvas evitaría este problema.

4. Transformaciones y reducción de dimensión

4.1 Introducción

Un paso opcional dentro de nuestro diagrama de flujo (Figura 4.1) son las posibles transformaciones u operaciones de reducción de dimensión que se pueden aplicar a los datos. Opcional en cuanto a que no tiene por qué ser fundamental para que la clasificación final tenga éxito, pero busca aumentar la capacidad discriminativa en el caso de las transformaciones o la eficiencia en los algoritmos de aprendizaje si hablamos de reducción de dimensión.

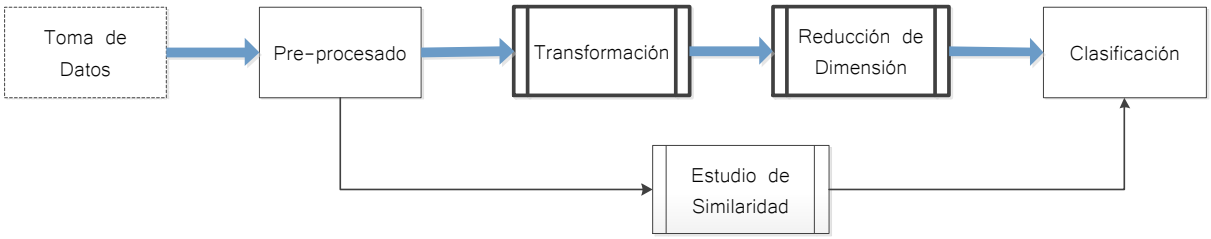


Figura 4.1. Diagrama de flujo.

4.2 Transformaciones

En esta sección se quiere dar cabida a dos cambios de representación de los datos que se utilizaron para resaltar ciertos aspectos de los datos o para facilitar su análisis.

4.2.1 Transformada Wavelet

El primero de ellos es la transformada Wavelet (TW), la cual realiza un análisis tanto en el tiempo (en el caso de señales temporales) como en la escala (vinculado a la frecuencia). Esta operación se basa en el producto e integración de la señal que se desea analizar por una función que sirve de ventana, denominada wavelet (Graps 1995).

Las funciones wavelet son funciones oscilatorias de duración finita. Se pueden obtener distintas versiones de ella modificando sus dos parámetros: la traslación y la escala. La traslación

controla la posición de la ventana y es la que proporciona la información temporal (o equivalente) de la señal. El parámetro de escala está relacionado a la información en frecuencia de la señal y controla el grado de detalle que se analiza. Al modificar la escala de la onda wavelet el efecto que tiene sobre ésta es su dilatación o compresión. El producto e integración de la transformada tendrá un valor distinto de cero en las partes de la señal en las que existen componentes de frecuencia similar a la que posee la onda wavelet. Los distintos valores que se obtienen para distintas combinaciones de traslación y escala conforman una transformada bidimensional de la señal original.

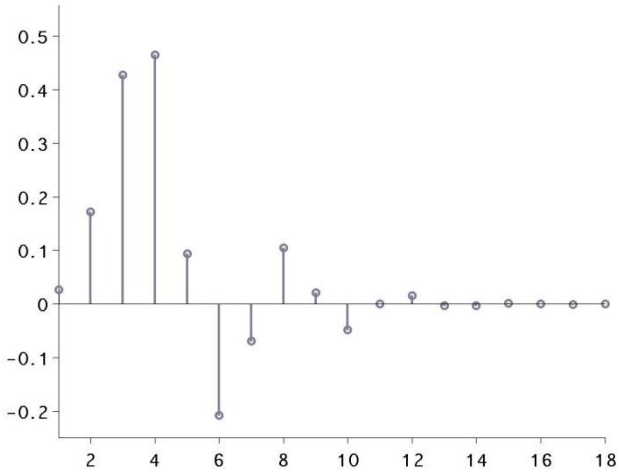


Figura 4.2. Función wavelet de la familia Daubechies tipo 9.

Una de las características de la TW es que para valores de escala pequeños se cuenta con una alta resolución temporal pero la resolución en frecuencia es baja, mientras lo contrario ocurre para escalas grandes.

Teniendo en cuenta que la TW contiene mucha información redundante, para nuestros propósitos es más eficiente la utilización de la transformada wavelet discreta (TWD). Ésta mantiene la idea de la TW, pero varía sobre todo en el método de cálculo y en el hecho de que la información de escala se obtiene para valores discretos de ésta.

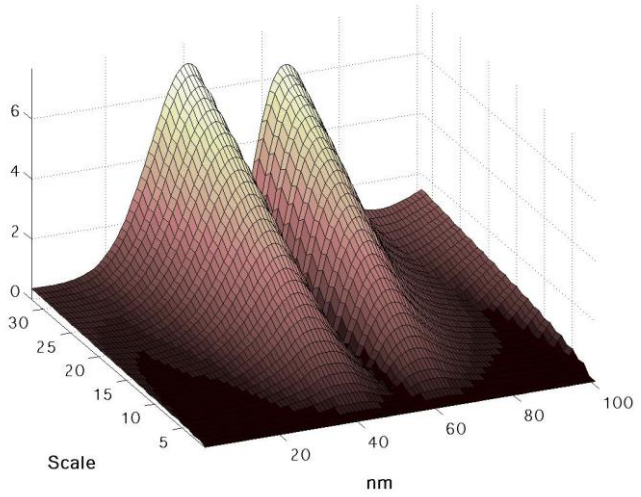
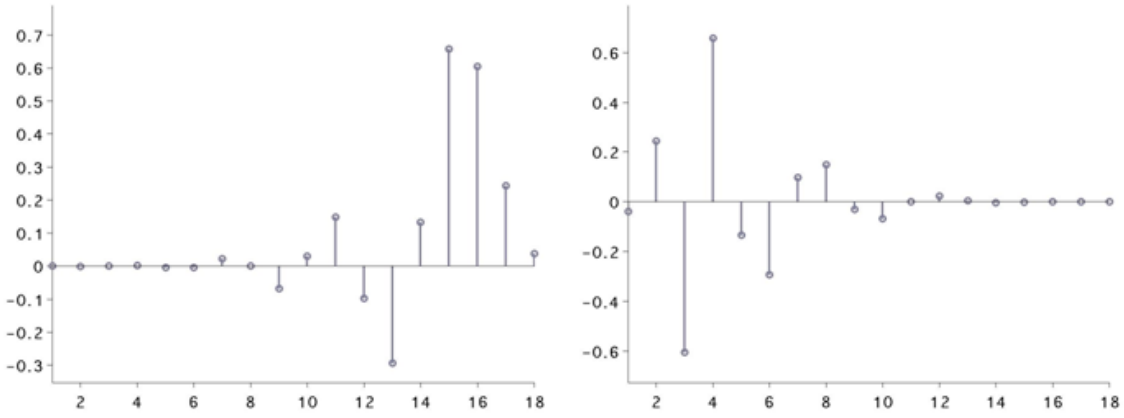


Figura 4.3. Representación del valor absoluto de la TW de una muestra de Thwi utilizando la función wavelet haar.

El método de cálculo consiste en filtrar la señal en sucesivos pasos para obtener la información contenida en distintos rangos de frecuencia. El proceso comienza filtrando la señal con un filtro paso alto y otro paso bajo. A la salida del primero se obtiene el primer nivel de coeficientes de la TWD. Del segundo se extrae la aproximación de la señal del primer nivel. Al haber reducido el ancho de banda a la mitad es posible un diezmado por dos. Los coeficientes diezmados son la entrada de la segunda pareja de filtros que dan como resultado el segundo nivel de detalles y de aproximación. En teoría se puede repetir el proceso hasta que la señal se reduce a una sola muestra. Las parejas de filtros se generan a partir de la función wavelet, una para la descomposición y otra para la reconstrucción.



(a) (b)

Figura 4.4. Ejemplos de filtros de descomposición paso bajo (a) y paso alto (b) de la función wavelet Daubechies 9.

La reconstrucción se realiza siguiendo los pasos inversos. En primer lugar se insertan ceros en los coeficientes del último nivel de detalles y de aproximación para doblar su tamaño al que tenían originalmente. A continuación se filtran a través de los filtros de reconstrucción y se suman. En el siguiente paso se hace lo propio con la señal reconstruida hasta el momento y los coeficientes de detalles del nivel inferior, y así hasta recuperar la señal original.

4.2.2 Análisis de la derivada

La mayor resolución de los sensores hiperespectrales ha permitido desarrollar y evaluar mejores métodos para el análisis de la forma espectral (Torrecilla et al. 2009). Uno de estos métodos es el análisis de la derivada, el cual consiste en el cálculo numérico de la derivada del espectro y a partir de ésta buscar nuevas características que hayan podido pasar desapercibidas en la curva original. Las funciones derivadas tienen la capacidad de resaltar cambios sutiles del espectro original que pueden ser importantes para la diferenciación.

Una de las formas de estimar la derivada es mediante el cálculo de las diferencias finitas divididas (Tsai & Philpot 1998):

$$\frac{ds}{d\lambda} = \frac{s[\lambda_i] - s[\lambda_j]}{\Delta\lambda} \qquad (4.1)$$

- $\Delta\lambda$ es la separación entre los puntos *j* e *i*.

En este caso se trataría de una diferencia progresiva puesto que para calcular la derivada en un punto solo se tiene en cuenta información posterior a él. Uno de los aspectos a tener en cuenta es la sensibilidad del cálculo de las diferencias finitas al ruido por lo que es vital realizar un suavizado previo.

La separación entre bandas no tiene por qué ser entre dos longitudes de onda. Si hacemos de $\Delta\lambda$ un parámetro seleccionable podremos ajustar el cálculo de la derivada al nivel de detalles que nos interese, efectuando además un suavizado adicional que hay que considerar (figura 4.5).

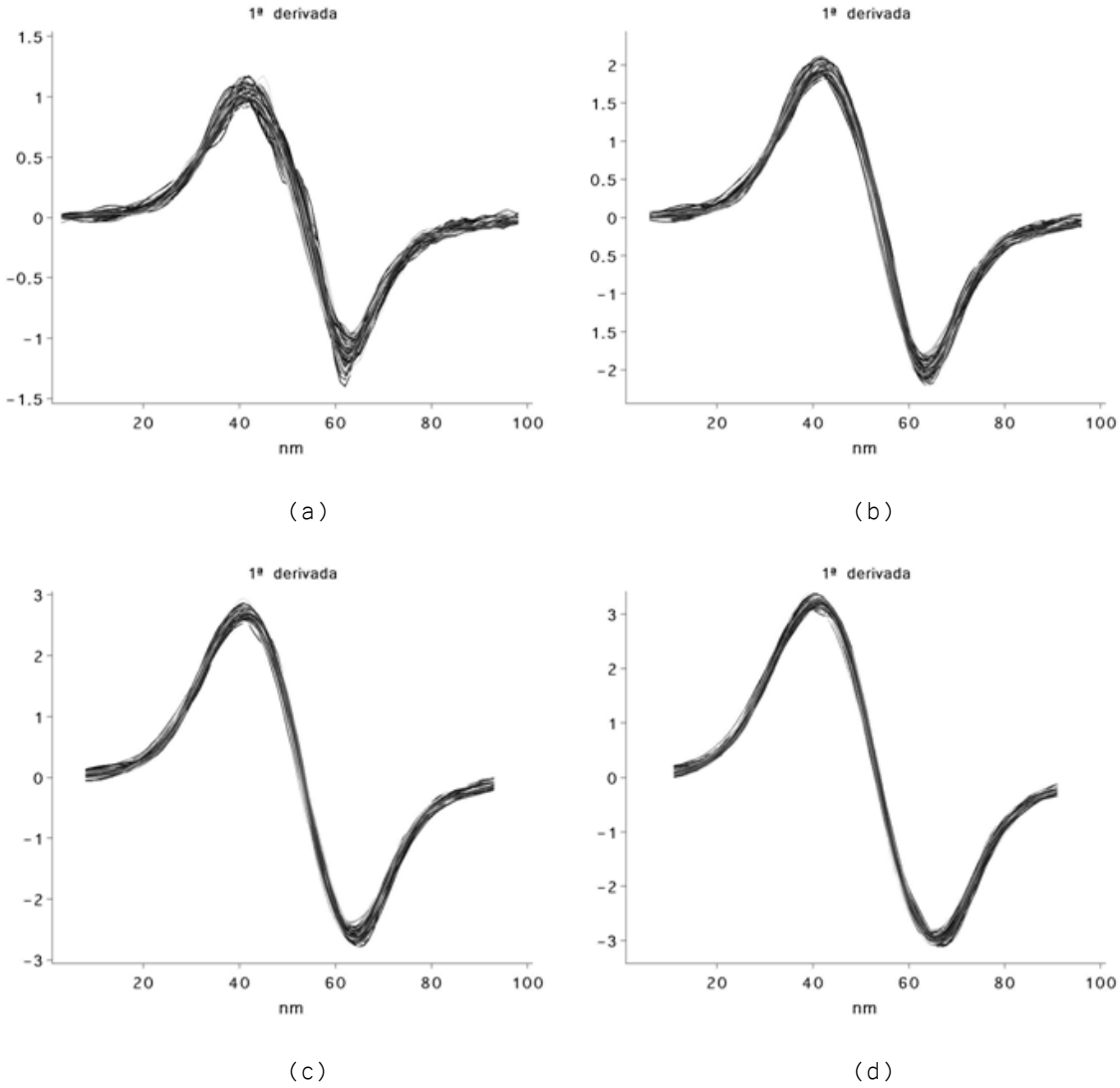


Figura 4.5. Primera derivada de las muestras de Thwi para distinta separación entre bandas. 5 (a), 10 (b), 15 (c), 20 (d). Se observa el efecto de suavizado que tiene utilizar una separación grande.

4.3 Reducción de la dimensión

La reducción de dimensión es el proceso de seleccionar u obtener un número de variables representativas de un conjunto de datos que originalmente tenían un número mayor de ellas. Por ejemplo cuando en el capítulo 2 se seleccionó el rango de longitudes con el que continuaríamos el estudio, se realizó una reducción de dimensión pasando de las 600 bandas disponibles a 101, dado que fuera de esta selección no esperamos encontrar información, en lo que a fluorescencia se refiere.

Idealmente la representación reducida de los datos debe tener una dimensionalidad igual a su dimensión intrínseca, ésta es el mínimo número de variables necesarias para representar los datos. Si se proyectara la información sobre menos variables estaríamos perdiendo información, mientras que si usáramos más estaríamos introduciendo variables ruidosas y redundantes (Levina & Bickel 2005).

La complejidad de la mayor parte de los algoritmos de aprendizaje depende de la dimensión de los datos de entrada, así como de la cantidad de muestras con la que se entrene. Para una menor ocupación de espacio en memoria y mejorar a su vez los tiempos de cómputo, es deseable reducir el número de variables. Así los clasificadores tienden a ser más simples y robustos, permitiendo una mejor generalización.

Una de las razones por las que una técnica de aprendizaje puede funcionar correctamente a pesar de manejar un gran número de variables es porque en realidad estos datos no tienen una alta dimensionalidad, sino que están incrustados en un espacio de alta dimensión. Ese podría ser el caso de nuestros datos, dado que su naturaleza como espectro hace que haya una alta dependencia entre las variables. Si escogemos una determinada longitud de onda, la fluorescencia medida un nanómetro por encima o por debajo, aportan relativa poca información adicional puesto tienen una fuerte correlación.

Un motivo práctico para la reducción de dimensión a tres o menos variables es la posibilidad de representar los datos en estos espacios. Un análisis visual de los datos puede revelar el solape de las variables de cada especie, cuáles de ellas son más semejantes y la estimación de la complejidad del modelo necesario para separar las clases.

Hay dos tipos principales de reducción de dimensión: selección de variables y extracción de variables (Alpaydin, p. 110). El primero consiste en elegir un número determinado de variables que conserven la mayor cantidad de información posible. En el segundo se crea un nuevo conjunto de datos de menor dimensión transformando los originales. La ventaja del primero es que permite una reducción del número de variables a medir, lo cual puede suponer menores tiempos de adquisición y menor coste. Por ejemplo si en lugar de tomar una medida hiperespectral, es suficiente con medir en unas pocas bandas entonces se podría utilizar instrumentos de menor coste con igual resultado. Sin embargo como para la extracción de variables se utilizan todas las variables originales, a igual número de ellas es más probable que acumulen más información de los datos que tras una selección.

Una de las formas más básicas de realizar la selección de variables es mediante la búsqueda del subconjunto con el menor número de dimensiones que mayor contribución hacen para discriminar entre las especies. Como una búsqueda exhaustiva es impracticable se ha optado por utilizar un tipo de algoritmo evolutivo para una realización más eficiente.

4.3.1 Algoritmo Genético

Inspirado en el proceso de la selección natural y la supervivencia del más fuerte, el algoritmo genético es un método heurístico para la resolución de problemas de optimización. El algoritmo modifica iterativamente una población de posibles soluciones para converger hacia una óptima, al menos desde un punto de vista local.

El proceso evolutivo comienza con una población de individuos llamados cromosomas generada aleatoriamente. En cada generación (iteración) se mide la aptitud de los individuos, es decir aquellos que están más cerca de la solución. Para formar la nueva generación se seleccionan aleatoriamente individuos llamados a ser los parientes para ser recombinados entre sí y crear nuevos cromosomas hijos, siendo más probable la selección de los individuos con mayor aptitud.

En primer lugar hay que seleccionar una forma de representación de los datos y una función de aptitud. Los individuos están representados mediante vectores cuyos genes (variables) representan una posible solución al problema. Los vectores pueden tener distintas codificaciones en función de la aplicación, por ejemplo la binaria puede ser utilizada para la selección de variables y la directa para hallar la solución de una función.

Por otro lado, la función de aptitud es aquella que se quiere optimizar. Si la aplicamos a los candidatos a solución podremos evaluar cuáles son los individuos que más cerca están de la solución y que por tanto conviene que su vector completo o parte de él se perpetúe durante las siguientes generaciones. En general la búsqueda continúa hasta que se alcanza un límite en el número de iteraciones o bien la aptitud medida para la población es satisfactoria.

Los hijos de la nueva generación pueden ser creados de tres formas distintas. La más habitual es la recombinación en la que se mezclan fragmentos de los vectores de dos parientes para formar uno nuevo. La segunda es la mutación en la que el individuo original modifica parte de su vector para tomar nuevos valores que puedan no estar representados en la población actual. Por último se puede permitir que aquellos individuos que hayan obtenido los mejores valores de aptitud sobrevivan en la siguiente generación. Éstos últimos formarían la llamada élite de la generación. El porcentaje que de cada uno de estos tipos de hijos se crean para la siguiente

generación es un parámetro de diseño. La selección de al menos un miembro de élite asegura que el mejor resultado de las sucesivas generaciones nunca sea peor que el de las anteriores.

En Vaiphasa et al. (2007) se aplica con éxito el algoritmo genético para seleccionar un subconjunto de bandas pertenecientes a espectros con mucha similitud inter-clase conservando una buena separabilidad entre ellas. Por tanto se planteó la posibilidad de utilizar el algoritmo para hacer lo propio con los espectros de emisión de fluorescencia, buscando la combinación de bandas que mejor conserva la información propia de cada especie y que por tanto permite seguir distinguiéndolas. Como buscamos la solución que mejores resultados para la clasificación automática de especies ofrece, la función de aptitud será precisamente la precisión que se obtiene con algún clasificador.

Uno de los factores que determinan la eficacia del algoritmo es la diversidad de la población. La diversidad es un indicativo de la variedad de los individuos en cada generación. Es importante que la población inicial estén representadas de forma uniforme todas las bandas del espectro para minimizar el riesgo de que se estén obviando zonas donde pueda encontrar la solución óptima. Sin embargo, si se mantiene esta diversidad en las sucesivas generaciones, el algoritmo puede no converger a una buena solución. La diversidad inicial se controla con el rango de valores que pueden tomar, mientras que la diversidad de las sucesivas generaciones depende del número de individuos que mutan y en qué grado lo hacen.

El tamaño de la población también es un parámetro importante. Cuanto mayor sea, más exhaustiva será la búsqueda y mejor será la solución que se obtiene, pero más largo será el proceso de búsqueda. El tiempo de cómputo en nuestro caso irá muy ligado al clasificador utilizado para evaluar la aptitud de las soluciones. Por ejemplo, un clasificador como el k-vecinos que no requiere entrenamiento permite una búsqueda más rápida que si se utilizara *self-organizing maps*, que requiere un proceso de aprendizaje (Capítulo 5).

Otro factor a tener en cuenta es la forma en la que se escalan los valores que proporciona la función de aptitud para distribuir la probabilidad de ser seleccionados a lo largo de la población. Un escalado muy restrictivo en el que solo se escogen a los mejores hará que éstos tomen rápidamente el control de la población, afectando de forma negativa a la diversidad.

El algoritmo se resume en las siguientes líneas:

- Se crea una población aleatoria en la que los individuos toman valores aleatorios.

- Se inicia la secuencia de sucesivas generaciones.
    - Se calcula la aptitud de cada individuo para evaluar lo buena que es su propuesta de solución.
    - Se escalan los valores de aptitud para facilitar la distribución probabilidades de selección a lo largo de toda la población.
    - Se seleccionan los parientes que crearán la siguiente generación.
    - Se crean los hijos de tres formas diferentes. Unos formarán parte de la élite, otros serán combinación de dos parientes y otros aparecerán como resultado de un cambio aleatorio en el vector de un pariente.
    - La nueva población se forma con los hijos creados.
- El algoritmo termina cuando se alcanza uno de los criterios de parada.

Antes de aplicar el algoritmo se utilizó un estimador de máxima verosimilitud (Levina & Bickel 2005) para estimar la dimensión intrínseca, cuyo resultado fue de cinco. Así pues utilizando el algoritmo genético se buscó la combinación de cinco bandas que mejores resultados de clasificación ofrece utilizando el algoritmo de k-vecinos con k = 2 (Capítulo 5).

Se decidió aplicar una codificación directa de forma que cada elemento de los vectores tomara cualquier valor entero en el rango de 1 a 101, correspondiente al número de bandas de los espectros. Después de unas cuantas ejecuciones preliminares del algoritmo se comprobó que las cinco bandas seleccionadas normalmente estaban escalonadas, cubriendo gran parte del espectro. Sin embargo un ajuste de los parámetros era necesario dado que el resultado final casi nunca se repetía. La configuración que finalmente dio resultados más estables fue la siguiente:

- Número de hijos élite: 1
- Porcentaje de recombinación: 0.6 (Mutación: 0.4)
- Probabilidad de mutar: 0.5 (Aplicado a cada elemento del vector de los parientes seleccionados)

Aparte de esto resultaba conveniente que los vectores tuvieran sus elementos ordenados de forma que al recombinar, independientemente de qué pariente provenga un elemento determinado, éste cubriera la misma zona del espectro. La mutación se realizó sumando o restando un número de bandas aleatorio extraído de una función de probabilidad normal para que cada componente realizara una búsqueda local.

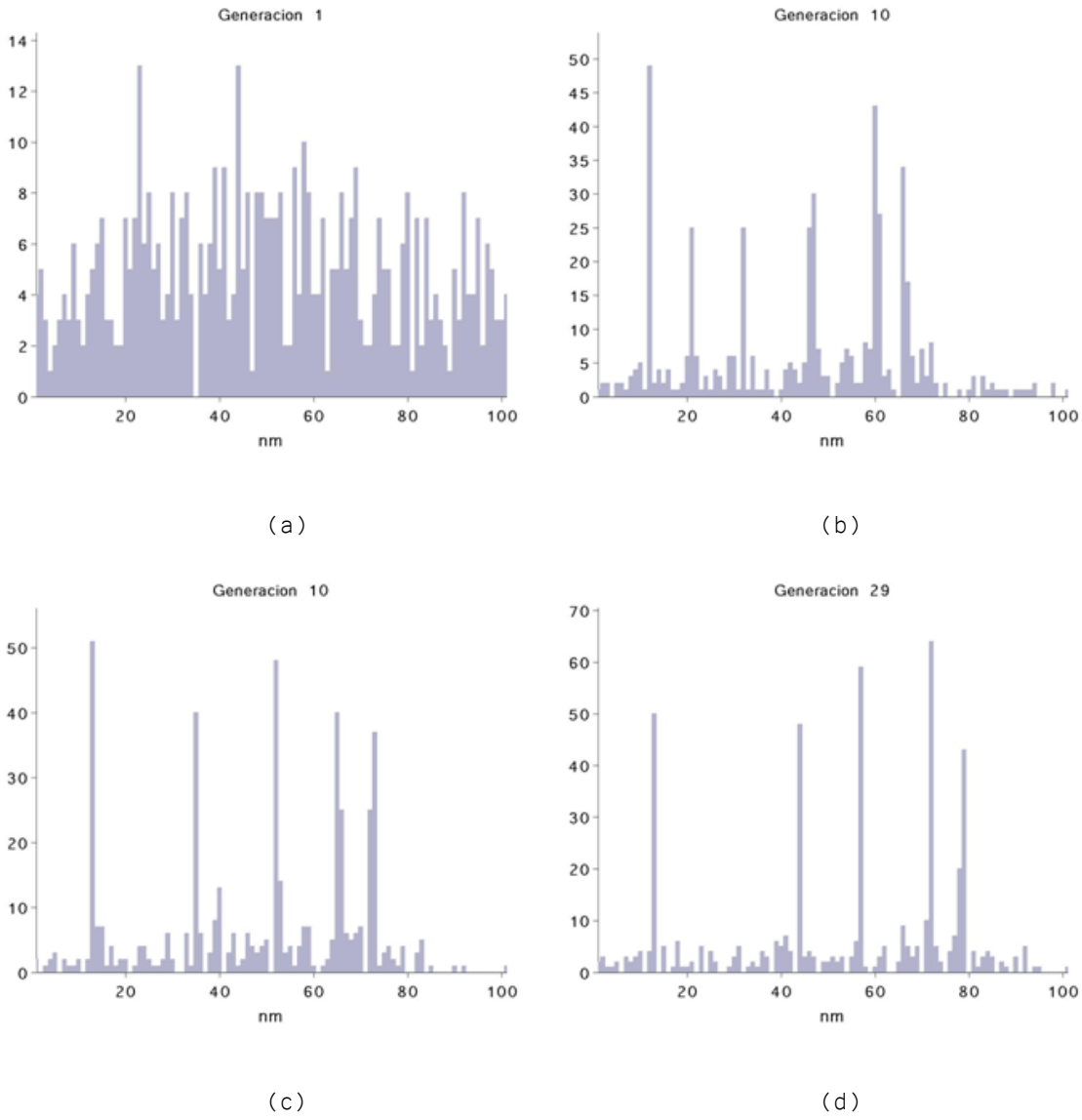


Figura 4.6. Presencia de las distintas bandas en la población de cuatro generaciones.

En la figura 4.6 se aprecia la evolución de las bandas que contiene la población de cromosomas (vectores). Inicialmente (a) la distribución debe ser lo más uniforme posible y a ser posible contener todas las bandas. Ya en la décima generación (b) destacan una serie de bandas como las más representativas hasta ese momento. En la última generación para este ejemplo (d) hay cinco bandas claramente destacadas, próximas a las que encontraría finalmente el algoritmo como óptimas. De las distintas realizaciones ejecutadas se infiere que las bandas en torno a las cuales se agrupan los resultados son:

$$651 - 666 - 677 - 688 - 694 \text{ [nm]}$$

El valor que retorna la función de aptitud para esta combinación de bandas es cero, es decir, clasificación sin errores, luego queda demostrada la viabilidad de aplicar la selección de variables sobre nuestros datos.

4.3.2 Principal Component Analysis

La técnica más extendida para realizar una extracción de variables es el análisis de componentes principales (Principal Component Analysis - PCA). Se trata de un método no supervisado de proyección en la que cada componente se calcula como una combinación lineal de las originales. El criterio a ser maximizado en el caso de PCA es la varianza.

La componente principal es la que al ser proyectados los datos sobre ella los datos están más dispersos de forma que la diferencia entre ellos se hace más evidente. Para que se cumpla esta condición de máxima varianza, la dirección de las proyecciones debe ser la definida por los vectores propios de la matriz de covarianza de los datos. Si ordenamos los vectores propios a partir de sus valores propios, la proyección sobre el vector con el mayor de ellos es el que maximiza la varianza y por ello es el primer componente principal. El vector con el segundo mayor valor propio es la segunda componente principal y la que más varianza de los datos contiene después de la primera. Procediendo de esta forma tendríamos ordenadas las proyecciones por importancia en cuanto a separabilidad de los datos. Escogiendo solo aquellos que abarcan la mayor parte de la varianza de los datos estaremos realizando una reducción de dimensión efectiva.

Si a partir de un determinado valor propio, sus valores propios son nulos entonces la dimensión efectiva de los datos es menor que la del espacio sobre el que se encuentran. Esto ocurre cuando la matriz de covarianza es singular. Si aun no siéndolo, su determinante es pequeño significa que hay valores propios con poca contribución a la varianza y pueden ser descartados.

La fórmula para el cálculo de las nuevas variables sería (Alpaydin, p. 114):

$$z = W^T(x - m) \qquad (4.2)$$

- Las columnas de *W* son los vectores propios del estimador de la matriz de covarianza de los datos ordenados por sus valores propios.

- x son los datos.
- m es la media de los datos y sirve para centrarlos.

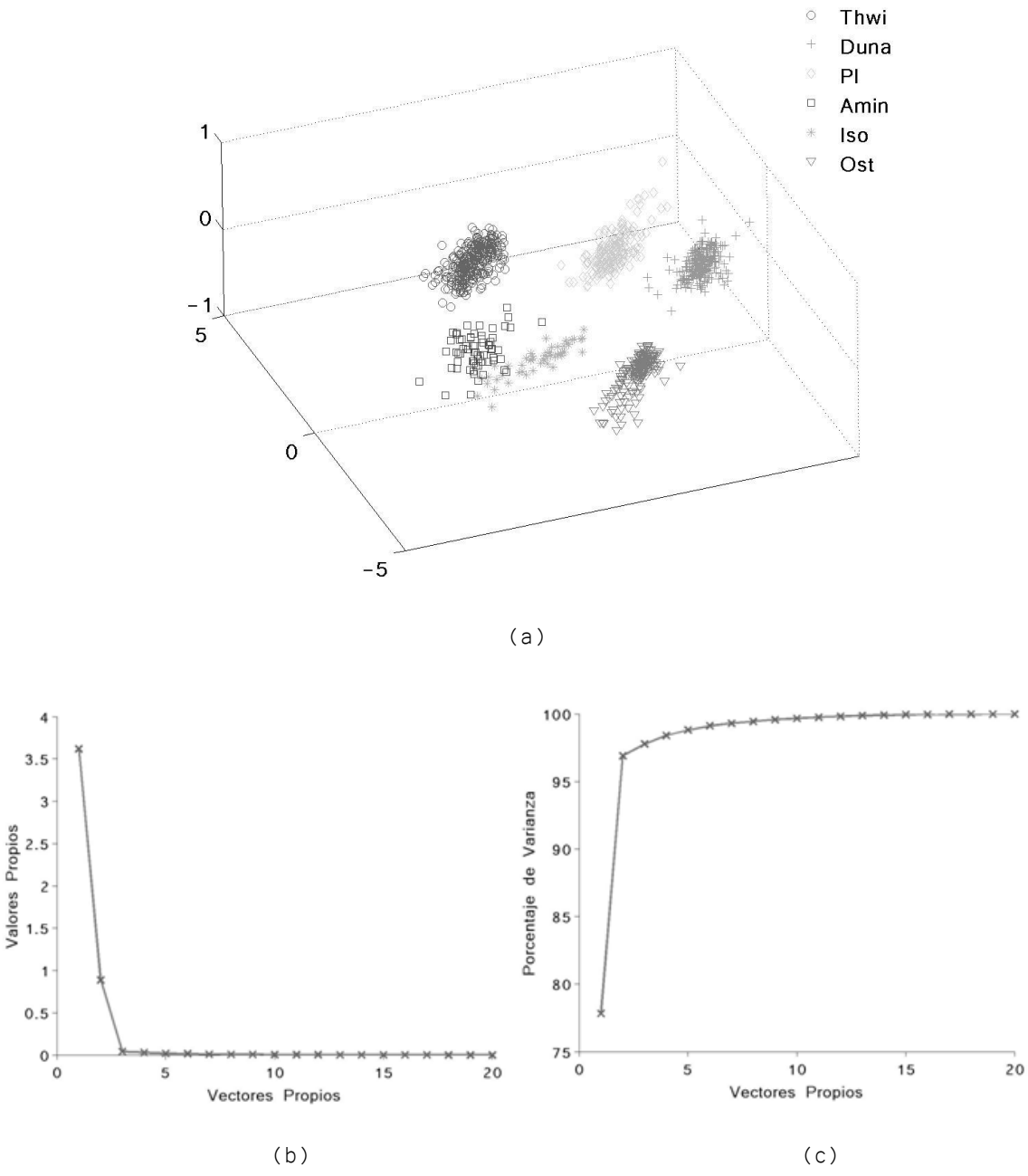


Figura 4.7. Representación de los datos de emisión de fluorescencia usando sus tres primeros valores propios (a). Los valores propios de los 20 primeros vectores propios (b) y el porcentaje de varianza contenido en ellos (c).

En la figura 4.7 vemos la drástica reducción en la dimensión que es posible realizar de los datos. Prácticamente las primeras dos principales componentes contienen la mayor parte de la

variabilidad de los datos (96.4%). Las gráficas de los valores propios y el porcentaje de varianza ayudan en la selección del número de componentes necesario para caracterizar los datos, descartando por ejemplo las variables que se encuentren a partir del codo.

Resumen

En el presente capítulo se han descrito dos transformaciones de los datos que se utilizaron durante el proyecto. En el caso de la TWD es la base una de las técnicas de suavizado y de una normalización empleada. En el caso de la derivada se utilizarán en el capítulo cinco para estudiar si proveen de información adicional que permita mejorar la distinción entre especies.

Una vez utilizadas dos formas distintas de reducir la dimensión de los datos queda claro que es posible conservar muy pocas variables sin que suponga apenas una pérdida de información, e incluso dando resultados de clasificación perfecta sobre el conjunto de validación (Capítulo 5) en el caso de la selección de variables con el algoritmo genético.

5. Similitud espectral entre especies

5.1 Introducción

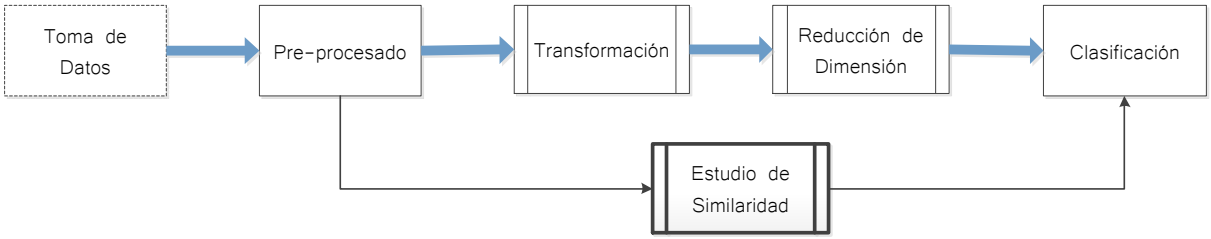


Figura 5.1. Diagrama de flujo.

Uno de los mayores problemas a los que se enfrenta un algoritmo de agrupamiento o clasificación a la hora de discriminar entre distintos patrones es el de encontrar una forma de medir lo similares o disimilares que son dos muestras de datos. En el caso que nos ocupa, tenemos un conjunto de muestras de espectros de emisión de fluorescencia pertenecientes a especies diferentes con una forma muy parecida entre ellos.

Lo que se pretende averiguar es si es apropiado con aplicar una medida de distancia habitualmente utilizada como es la euclidiana, o si proceden otros métodos como podría ser la distancia angular. Una buena medida de similitud sería aquella que lograra que los espectros de una especie tuvieran un índice de similitud grande (o una distancia pequeña intra-especie) a la par que manteniendo lo más alejado posible a los espectros del resto de las especies mediante un índice de similitud pequeño (o una distancia grande inter-especie).

En este capítulo se presentan las medidas de distancia o índices de similitud habitualmente usados, como son las geométricas. Luego se presentan algunas basadas en la codificación de los espectros para luego pasar a aquellas basadas en la estadística. Sin embargo, primero se necesita una forma de evaluar el desempeño de las medidas de similitud. Con este propósito se presentan primero de todo algunos índices de discriminabilidad utilizados en el campo de la teledetección.

5.2 Índices de discriminación

Cuando se cuenta con una serie de espectros asignados a distintas clases y se tiene una muestra que se desea identificar, resulta de interés obtener algún parámetro que proporcione información objetiva sobre la capacidad de una determinada medida de similitud para discriminar entre las diferentes clases, es decir, la probabilidad de que esa muestra pueda ser identificada convenientemente.

Gran parte de la literatura se ha volcado en estudiar el problema para el caso de imágenes de teledetección hiperespectrales (Chang 2000, Kong et al. 2010) Algunas de las técnicas usadas en este campo son extrapolables al caso general de la espectroscopia debido a que cada uno de los pixeles que conforman la imagen es considerado como un vector con una dimensión igual al número de bandas espectrales, formando así un espectro continuo.

Con el fin de aclarar conceptos, en este documento una medida de discriminabilidad espectral es un parámetro que sirve para evaluar los resultados que ofrece una medida de distancia o índice de similitud. El escenario en el que se van a aplicar es el siguiente: contamos con una librería de espectros digamos "estándar", uno por cada especie disponible, y con un conjunto de muestras de cada una de ellas que usaremos para medir distancias con los espectros de la librería. Si queremos hacer una clasificación simple calcularemos la distancia entre cada muestra con las referencias y le asignaremos la especie para la que se haya obtenido un resultado menor. Esta forma de clasificar no deja de ser igual a la técnica de los k vecinos con k igual a 1 y usando una muestra de entrenamiento por clase (capítulo 6).

Según cuál sea la librería de espectros de referencia utilizado los resultados de los parámetros cambiarán, por tanto es importante hacer una selección lo más representativa posible. Como vimos en el capítulo de pre-procesado (capítulo 3), cuando los datos de una especie están normalizados, éstos tienden a agruparse en torno a un cierto valor. Estadísticamente, cuanto más cercana pase una curva por los valores intermedios de cada longitud de onda, más probabilidades habrá de que pertenezcan a la misma especie. Por tanto la forma más sencilla de elaborar esta librería es usando las medias de las muestras de cada especie. Volviendo a la analogía con k vecinos, si tuviéramos que condensar los datos para quedarnos solo con uno por clase, seguramente la decisión más acertada sería quedarnos con la media muestral.

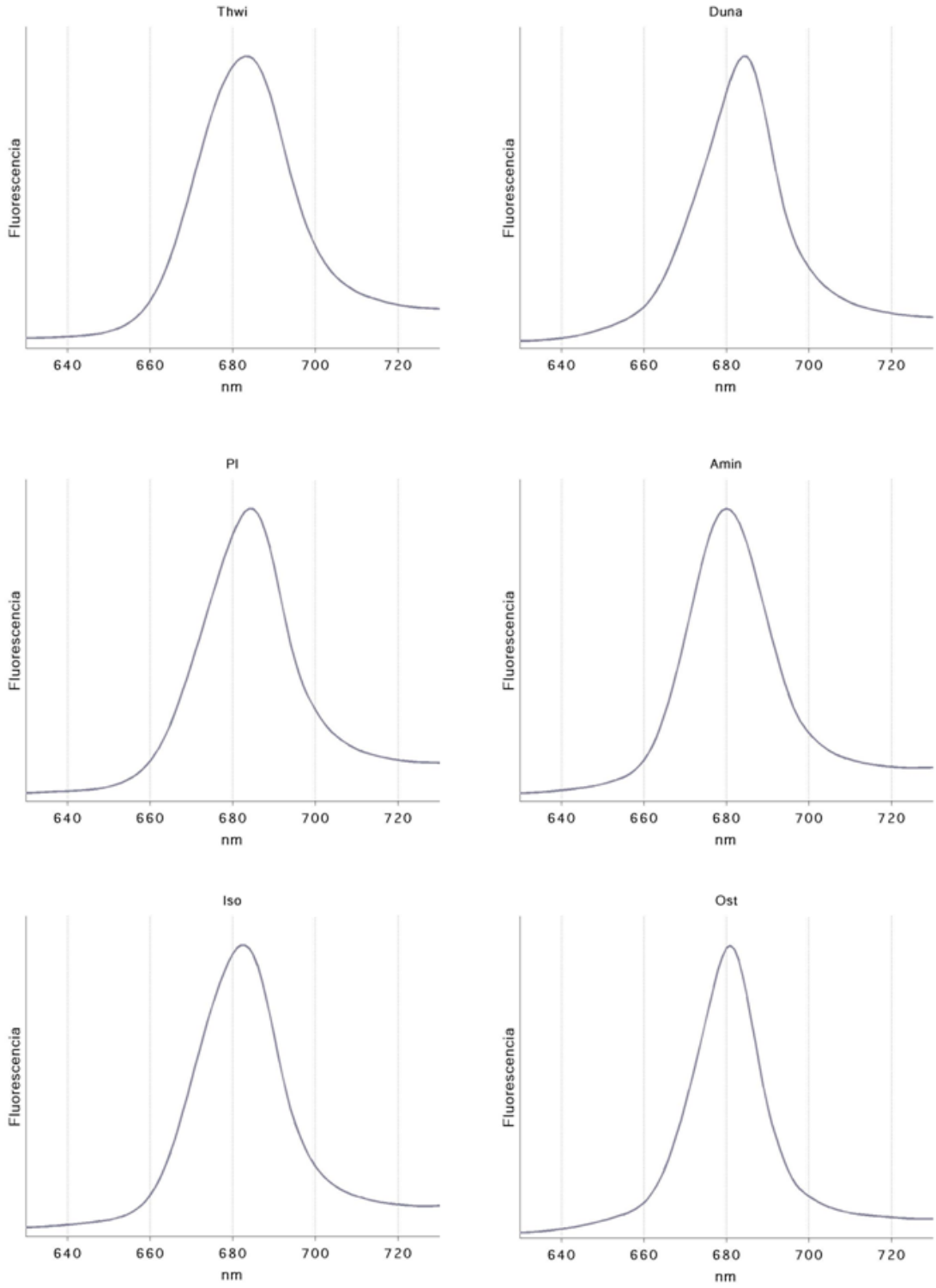


Figura 5.2. Espectros de referencia de cada especie calculados a como la media de las muestras una vez aplicada la normalización basada en wavelet y el suavizado combinado de wavelet y Satizky-Golay.

Para elaborar la librería de espectros de referencia de la figura 5.2 se utilizó la normalización basada en wavelet y el suavizado combinado de Wavelet y Savitzky-Golay, presentados en el capítulo 3.

Si pretendemos que un programa de ordenador aprenda a distinguir especies usando unas formas tan parecidas primero debemos entender cómo las distinguimos nosotros. Aprovechando que tenemos un conjunto reducido de muestras representativas de cada especie vamos a analizar los matices que las diferencian.

La característica más evidente es el ancho del pico. Thwi posee el más ancho, mientras que Duna y Ost ostentan el más estrecho. Otra podría ser la inclinación del espectro. Mientras en algunos casos, como el de Pl, parece tener una ligera inclinación hacia la derecha, en otros, como el de Amin, la desviación es hacia la izquierda. La posición del máximo también aporta información. Por ejemplo Duna y Ost tienen sus máximos separados por una media de 4 nm. Las pendientes también pueden ser útiles dado que como se aprecia, las faldas de los espectros terminan con alturas diferentes.

5.2.1 Probabilidad de Discriminación Relativa (PDR)

Volviendo al propósito de este capítulo, el primer índice de discriminabilidad que se va a utilizar es la Probabilidad de Discriminación Espectral Relativa (RSDPB). Ésta es una medida muy simple en la que se tiene en cuenta cada una de las distancias que una muestra desconocida tiene respecto a las de referencia. La RSDPB de cada especie de la librería se calcula como la distancia que tiene la muestra con la referencia dividido por todas las demás distancias. Como su nombre indica es una medida relativa puesto que se tiene en cuenta la distancia con el espectro de la especie correcta como la del resto, por lo que para tener un buen índice no solo importa que la distancia intraclase sea pequeña sino que sea grande para las inter-clase. Cuanta menor sea esta probabilidad asociada a un espectro de referencia, más cercana está la muestra a ella respecto a los demás.

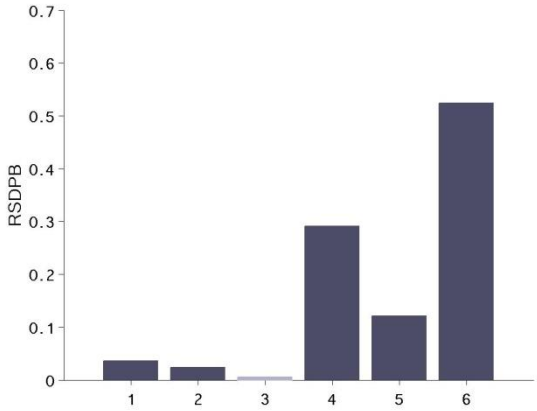


Figura 5.3. Un ejemplo de RSDPB en el que el espectro de la muestra bajo estudio pertenece a Pl (clase número 3). Al tener el valor más pequeño de RSDPB, se puede afirmar que en general las muestras de esta especie están más cerca de su espectro de referencia que de las de cualquier otra. La especie número 6 (Ost) aparece como la más distante respecto de las muestras de Pl.

Uno de los problemas de RSDPB es el que se pone de manifiesto en la figura 5.3. Al haber una distancia muy grande con la clase número 6 (Ost), su probabilidad tiene un valor muy alto comparado con el resto y aunque se aprecia que la de Pl es la menor es difícil saber si hay mucha diferencia con el que le sigue, Duna. Otro inconveniente es el hecho de que el espectro de referencia con menor probabilidad de discriminación sea en realidad el que más probabilidades tiene de compartir especie con la muestra bajo comparación no es demasiado intuitivo.

Por ello se propone utilizar una modificación de la RSDPB a la que nos referiremos simplemente como Probabilidad de Discriminación Relativa (PDR). Consiste básicamente en calcular las probabilidades sobre las distancias invertidas, de forma que las distancias muy pequeñas sean las que destaquen sobre el resto. La fórmula de las probabilidades quedaría de la siguiente forma:

$$RSDPB(i) = \frac{d(t,s_i)}{\sum_{j=1}^{N} d(t,s_j)}, \quad para\ i = 1,2,..,N \qquad (5.1)$$

$$RSDPB(i) = \frac{\frac{1}{d(t,s_i)}}{\sum_{j=1}^{N} \frac{1}{d(t,s_j)}}, \quad para\ i = 1,2,..,N \qquad (5.2)$$

La nomenclatura es la siguiente: *d* es la distancia utilizada, *t* es la muestra bajo estudio, *s* es el conjunto de espectros de la librería y *N* el número de especies contenidas en ellas. El resultado es ahora más fácil de interpretar (figura 5.4).

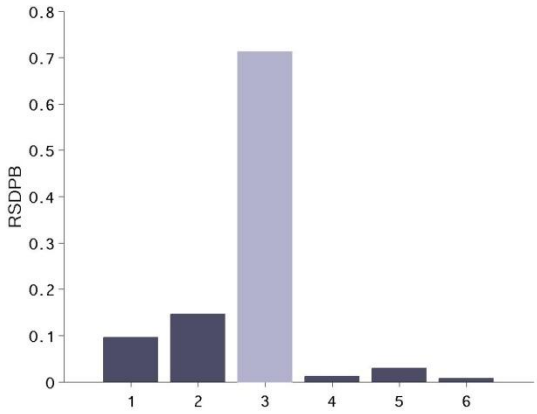


Figura 5.4. Un ejemplo de Probabilidades de discriminación relativa (PDR) en el que el espectro la muestra bajo estudio pertenece a PI (clase número 3).

5.2.2 Entropía de Discriminación Relativa (EDR)

Otro índice que aparece en la literatura es la Entropía de Discriminación Espectral Relativa (RSDE). Ésta se calcula como la entropía de las probabilidades de RSDPB, por lo que hereda los problemas que de éste índice apuntamos. Como expone Chang (2003) una RSDE es solo indicativo de que hay un espectro de la librería muy distante a la muestra bajo estudio, mientras que el resto de distancias puede ser todo lo pequeñas que quieran. El objetivo sería un parámetro cuyo valor fuera indicativo de lo fácil que podría ser clasificar la muestra, lo cual sucede si la distancia con el espectro de la librería correcto es pequeña y la del resto indefinidamente grande respecto a ésta.

Si en lugar de RSDE calculamos la entropía sobre la PDR entonces sí encontramos un parámetro adecuado para nuestros propósitos al que nos referiremos como la Entropía de Discriminación Relativa (EDR). Cuánto más cercana a cero sea la EDR mayor discriminabilidad proporciona la distancia en cuestión, mientras que el peor caso sería cuando existe la misma probabilidad para todas las clases y por tanto el valor de la entropía sería el de su cota superior, es decir, el logaritmo del número de clases, en nuestro caso seis.

$$EDR = -\sum_{j=1}^{N} PDR(j) \cdot \log PDR(j) \qquad (5.3)$$

5.3 Medidas de distancia

En esta sección pasamos a describir las medidas de distancia que serán evaluadas. Las primeras presentaremos son las geométricas, y dentro de éstas la medida de distancia euclidiana (EDM). En el terreno de las comunicaciones, el procesado de señales y el reconocimiento de patrones, este tipo de medida es el más común junto a las angulares. Los principales inconvenientes de la distancia euclidiana son por un lado su sensibilidad a la escala de las variables implicadas y por otro su incapacidad para detectar la correlación existente entre las variables o bandas que conforman el vector de información hiperespectral. Como consecuencia de esto, dos espectros con una misma forma espectral pero tomadas bajo configuraciones instrumentales distintas tendrán una distancia geométrica entre ellos elevada debido a que sus intensidades relativas diferirán. Por tanto se hace imprescindible normalizar los datos antes de lanzarse a calcular distancias euclidianas

La expresión generalizada de las medidas de distancia es la llamada distancia de Minkowski:

$$\mathrm{MD}(x,y) = \left(\sum_{i=1}^{N} |x_i - y_i|^p\right)^{1/p} \qquad (5.4)$$

Los valores de p más usuales son $p = 1$, $p = 2$ y $p = \infty$. Si $p = 1$ se obtiene la distancia Manhattan (City Block Distance):

$$\mathrm{CBD}(x,y) = \sum_{i=1}^{N} |x_i - y_j| \qquad (5.5)$$

Hacer tender p a infinito da lugar a la distancia Tchebyshev:

$$\mathrm{TD}(x,y) = \max_{i=1}^{N} \{|x_i - y_j|\} \qquad (5.6)$$

Sin embargo lo que utilizaremos es $p = 2$ se trata de la distancia más habitual e intuitiva, la euclidiana:

$$\mathbf{ED}(x,y) = \left[\sum_{i=1}^{N}(x_i - y_j)^2\right]^{1/2} \quad (5.7)$$

Otra distancia habitualmente utilizada es la de Mahalanobis en la que interviene la matriz de covarianza para tener en cuenta la correlación entre las variables para ponderar la separación entre ellas.

$$\mathrm{MD}(x,y) = [(x-y)^T S^{-1}(x-y)]^{1/2} \quad (5.8)$$

Sin embargo debido a que para algunas de las especies la dimensión es mayor que el número de muestras disponibles, la matriz de covarianza es singular y por tanto no invertible (Ledoit & Wolf 2002).

La siguiente distancia que usaremos es la Medida Angular Espectral (SAM), la cual mide la similitud entre dos muestras hallando el ángulo formado por los vectores que definen sus espectros. Dada la naturaleza invariante del ángulo a variaciones lineales de escala, cuando dos espectros difieren tan solo en un factor de escala m el ángulo que forman será nulo mientras que la distancia euclidiana sería proporcional a m. Su expresión viene dada por:

$$\mathbf{SAM}(x,y) = \cos^{-1}\left(\frac{\sum_{i=1}^{N} x_i y_i}{\left(\sum_{i=1}^{N} {x_i}^2\right)^{1/2}\left(\sum_{i=1}^{L} {y_i}^2\right)^{1/2}}\right) \quad (5.9)$$

Si tanto *x* como *y* están normalizados a la unidad, es decir, cada componente se divide por la suma de todos ellos, es posible establecer una relación entre ED y SAM a través de la siguiente equivalencia (Van der Meer 2006):

$$ED(x,y) = 2sin\left(SAM(x_i,y_j)/2\right) \quad (5.10)$$

Dado que el sin(x) y x son infinitésimos equivalentes, si la medida de distancia angular es pequeña entonces es equivalente a la euclidiana.

El coeficiente de correlación (Robila & Gershman 2005) proporciona una medida de similitud robusta, menos sensible al ruido que la distancia euclidiana (Van Der Meer & Bakker 1997) e incluso apta para evaluar distancias entre espectros sin normalizar. Se obtiene mediante la siguiente expresión:

$$\mathbf{CC}(x,y) = \cos^{-1}\left(\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{N}(x_i-\bar{x})^2\right]^{1/2}\left[\sum_{i=1}^{L}(y_i-\overline{y})^2\right]^{1/2}}\right) \qquad (5.11)$$

Las medidas de similitud basados en codificación (Du et al. 2003) se utilizaron para estudiar si a pesar de la pérdida de información que se produce al utilizar estas técnicas todavía seguía siendo posible distinguir entre especies. Estos métodos fueron concebidos para permitir un rápido procesado de los datos, especialmente teniendo en cuenta el volumen de datos que se maneja en el ámbito de la teledetección.

La codificación binaria consiste en dar a cada componente del espectro el valor 1 si se encuentra por encima de un umbral o 0 en caso contrario. El umbral puede ser cualquiera pero lo usual es utilizar la media. La distancia entre dos muestras codificadas de esta forma consistiría en contabilizar el número de componentes cuyos valores difieren.

La codificación cuaternaria en lugar de utilizar un umbral establece tres, dividiendo el espectro en cuatro zonas. Los segmentos del espectro pertenecientes a cada una de las zonas reciben una codificación distinta. Para el primer umbral también se utiliza la media como en el caso binario, y las medias de cada una de las dos zonas que divide fijan los otros dos.

Por último, aparte de medir las distancias sobre los espectros originales, también se probará sobre su derivada para conocer más sobre su capacidad discriminativa.

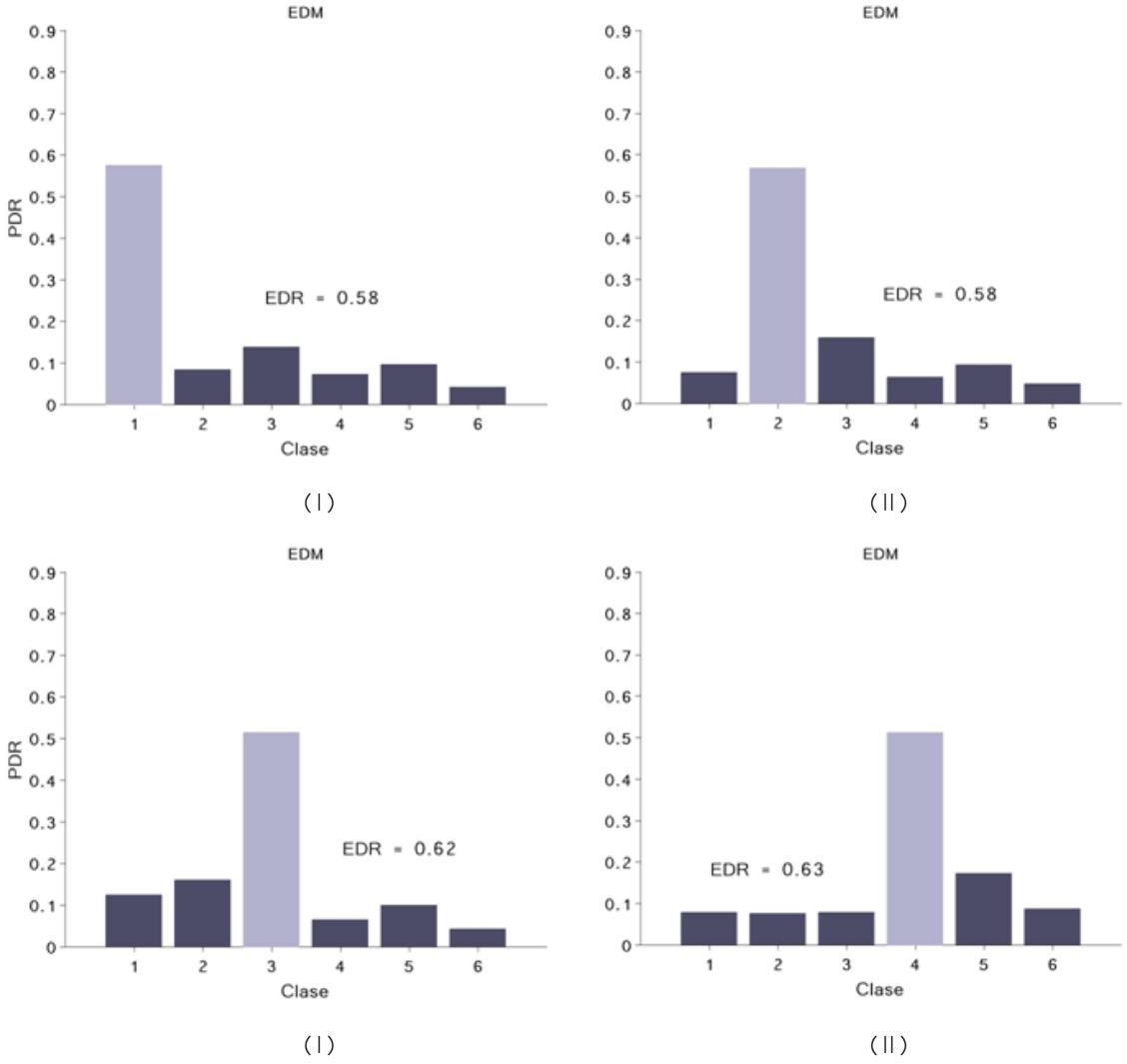
5.4 Metodología y resultados

El objetivo será utilizar los índices de discriminabilidad para evaluar qué medida de distancia ofrece mayor separabilidad entre los datos. La librería de espectros de referencia de cada especie estará formada por el espectro medio de cada una de ellas por lo que se contará con seis. Se midieron las distancias de todas las muestras disponibles de la primera toma (tabla 5.1) y para cada especie se calculó la media de las distancias con cada espectro de referencia, quedando una matriz de 6x6. Por ejemplo el elemento 1,2 de la misma es la media de las distancias de las muestras de la clase 1 (Thwi) con el espectro de referencia de la clase 2 (Duna). Por cada fila de esta matriz se extraen la PDR y la EDR que serán los resultados presentados.

Nuevamente se ha utilizado la normalización de wavelet y para el suavizado también el basado en wavelet combinado con Savitzky-Golay.

Especie	Nº de muestras	Nº de clase
Thwi	300	1
Duna	249	2
PI	201	3
Amin	65	4
Iso	52	5
Ost	180	6

Tabla 5.1. Información sobre el número de muestras total por cada clase bajo estudio.



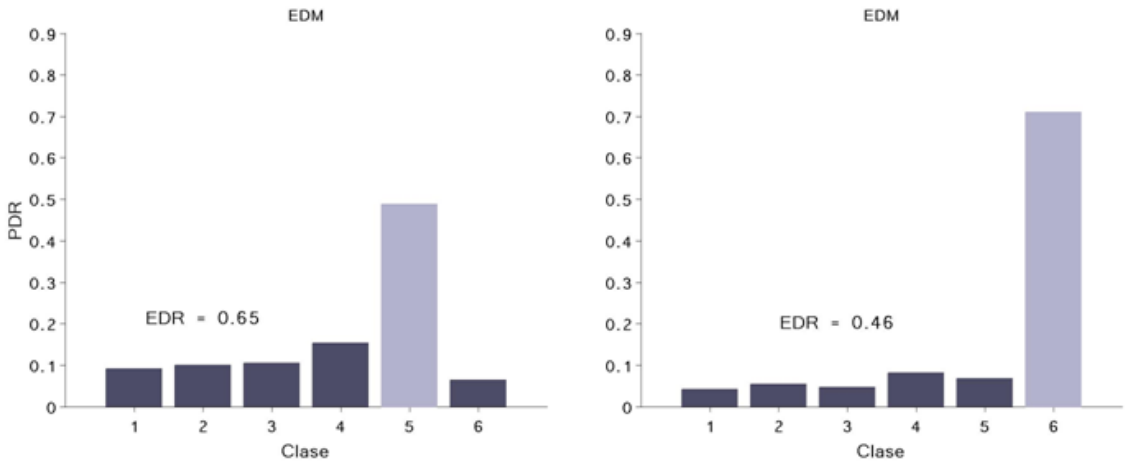
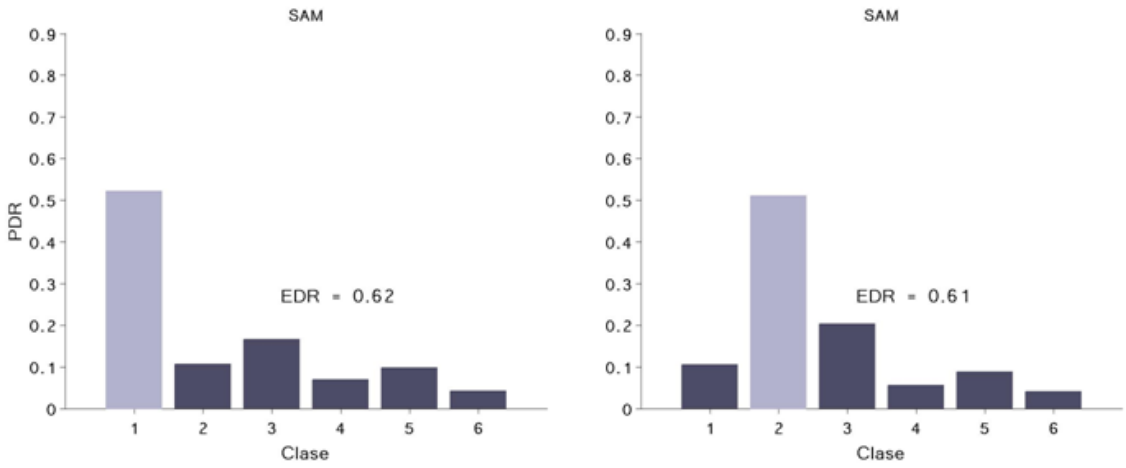


Figura 5.5. Cálculo de las PDRs y las EDRs usando distancia euclidiana (EDM). En cada gráfica la barra en gris claro representa la clase de las muestras comparadas con la librería. Una PDR grande significa una mayor cercanía a las muestras utilizadas respecto al resto de las clases. Cuanto menor es la EDR, más fácil es la discriminación porque una clase destaca sobre todas las demás por su cercanía a las muestras.



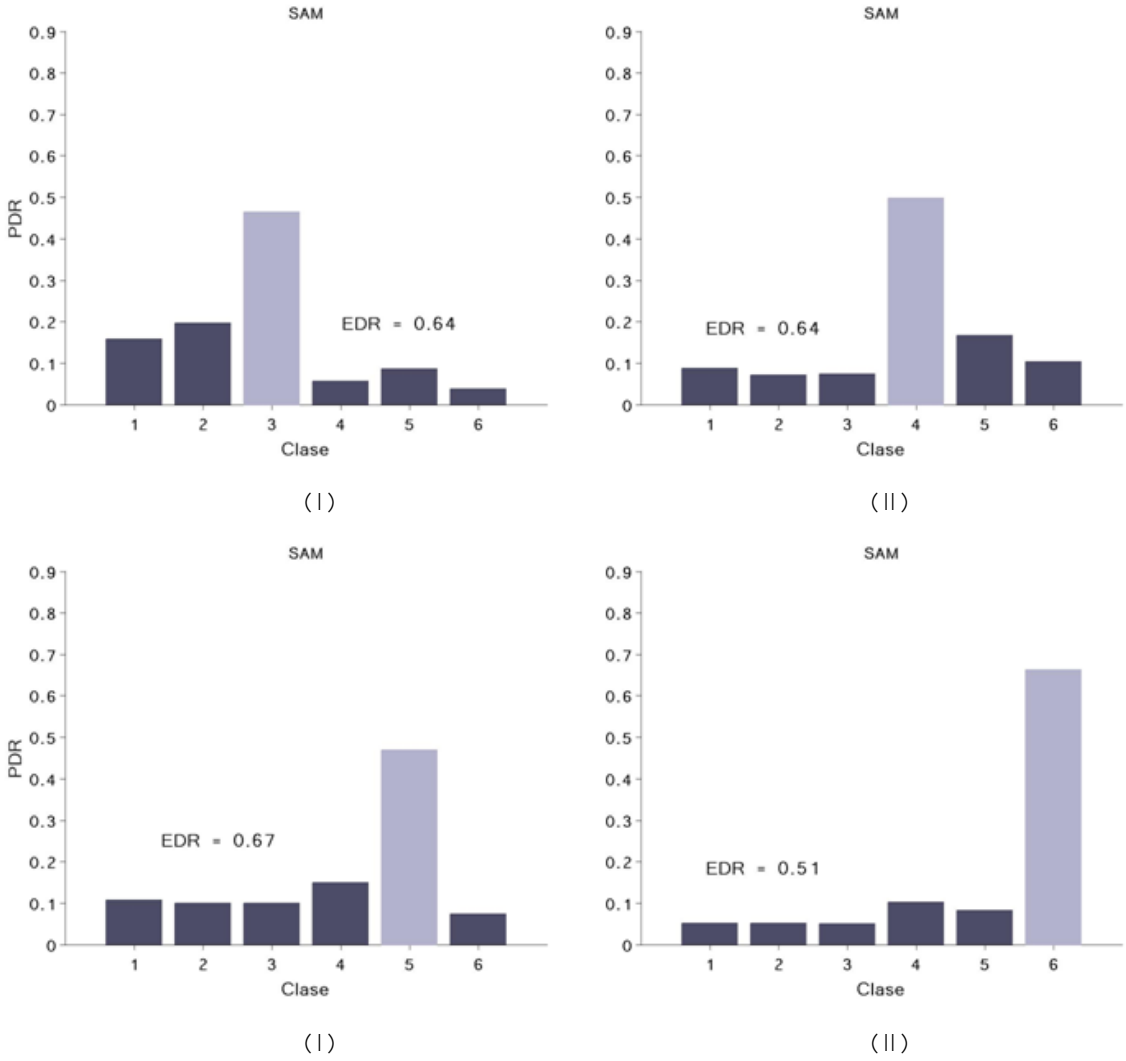


Figura 5.6. Cálculo de las PDRs y las EDRs usando la medida angular espectral (SAM). En cada gráfica la barra en gris claro representa la clase de las muestras comparadas con la librería. Una PDR grande significa una mayor cercanía a las muestras utilizadas respecto al resto de las clases. Cuanto menor es la EDR, más fácil es la discriminación porque una clase destaca sobre todas las demás por su cercanía a las muestras.

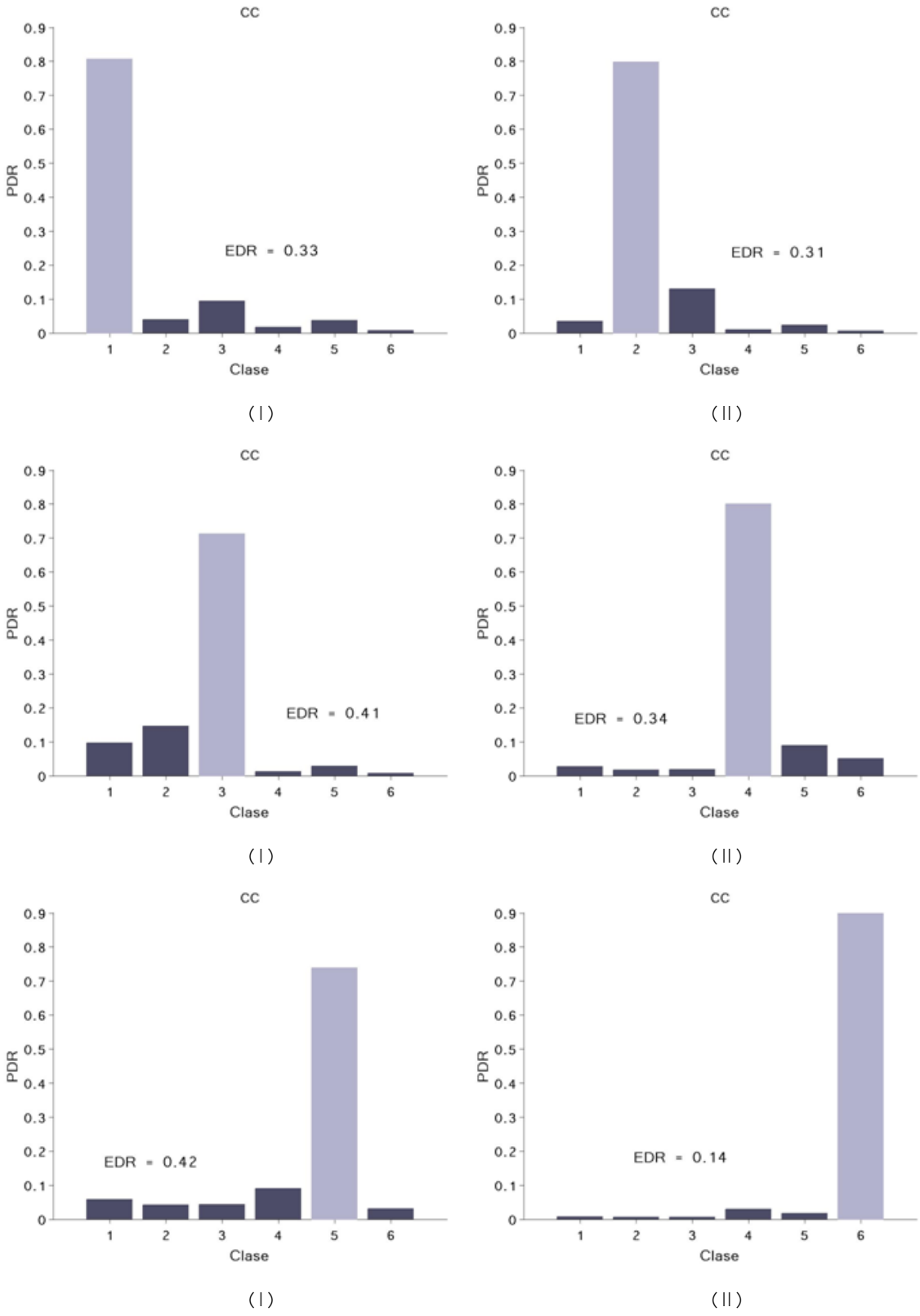
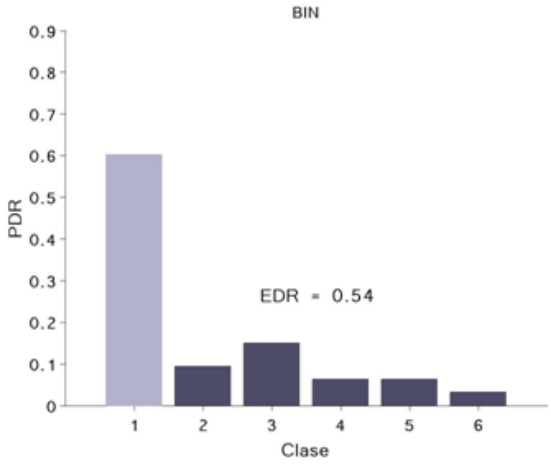
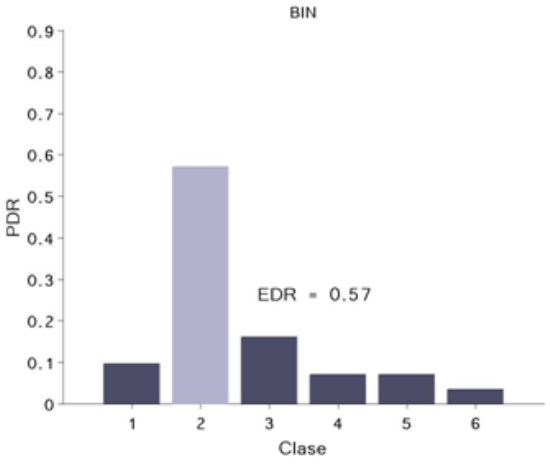


Figura 5.7. Cálculo de las PDRs y las EDRs usando el coeficiente de correlación (CC). En cada gráfica la barra en gris claro representa la clase de las muestras comparadas con la librería. Una PDR grande significa una mayor cercanía a las muestras utilizadas respecto al resto de las clases.

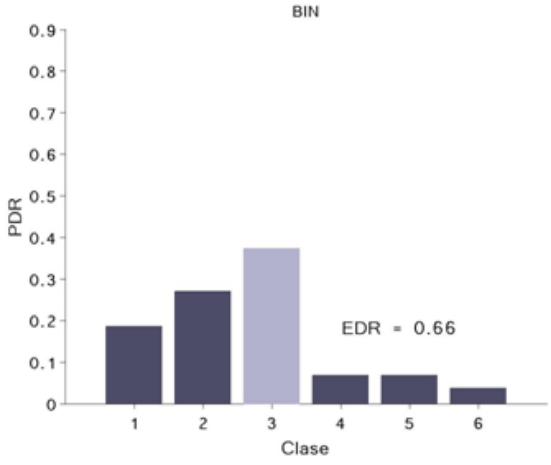
Cuanto menor es la EDR, más fácil es la discriminación porque una clase destaca sobre todas las demás por su cercanía a las muestras.



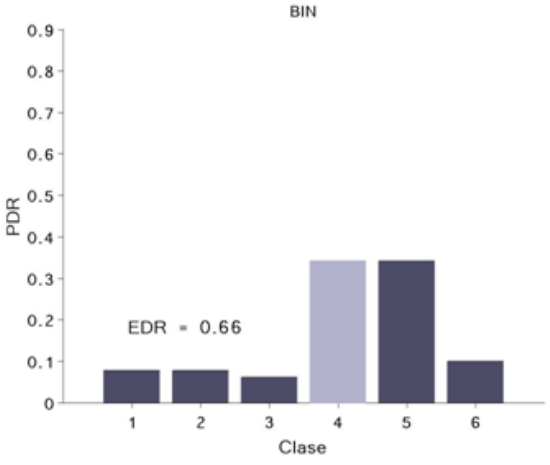
(I)



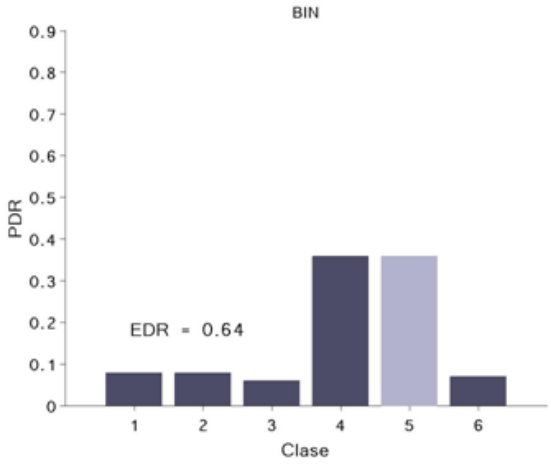
(II)

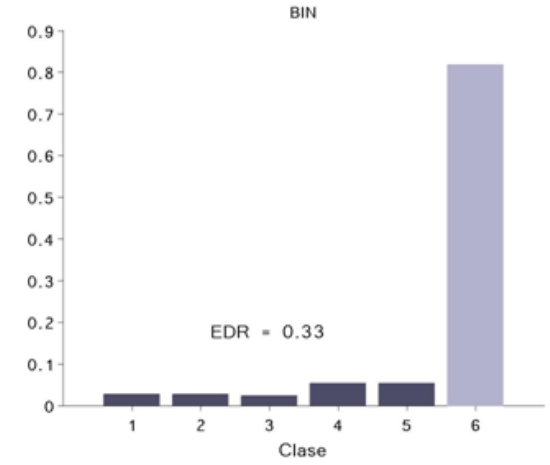


(I)



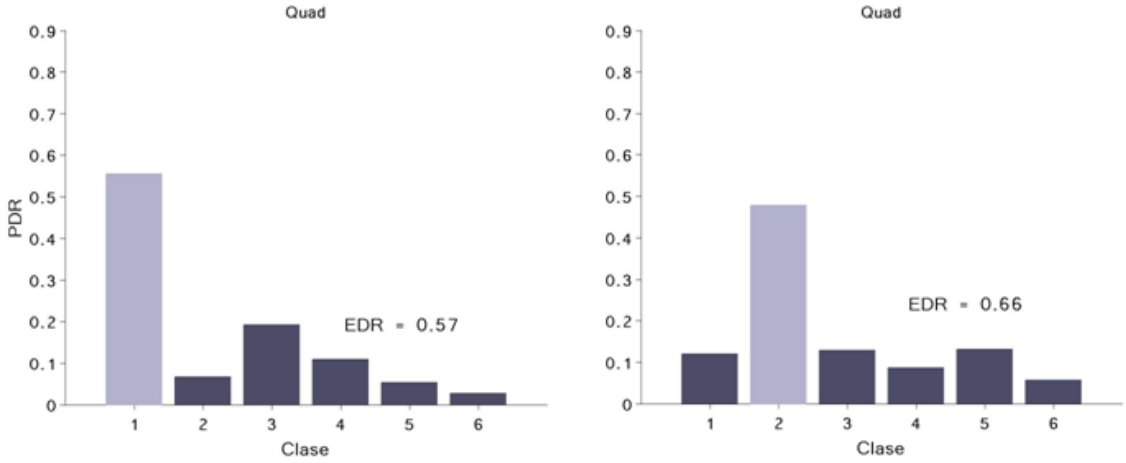
(II)



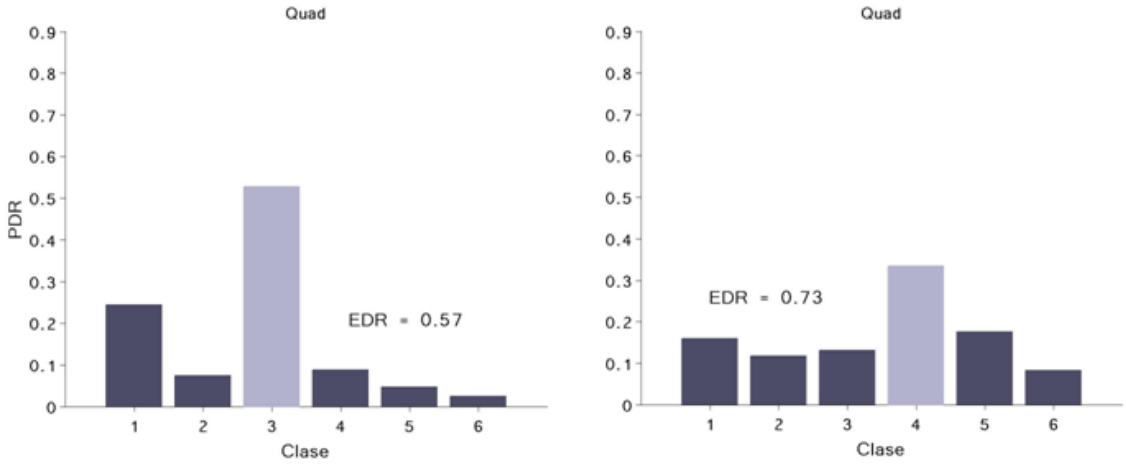


(I) (II)

Figura 5.8. Cálculo de las PDRs y las EDRs usando el codificación binaria. En cada gráfica la barra en gris claro representa la clase de las muestras comparadas con la librería. Una PDR grande significa una mayor cercanía a las muestras utilizadas respecto al resto de las clases. Cuanto menor es la EDR, más fácil es la discriminación porque una clase destaca sobre todas las demás por su cercanía a las muestras.



(I) (II)



(I) (II)

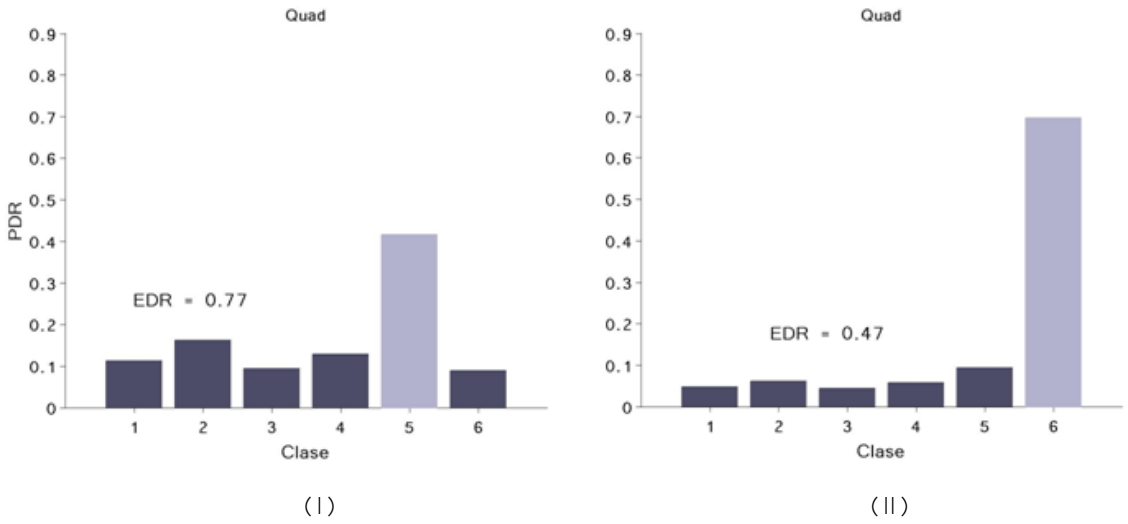
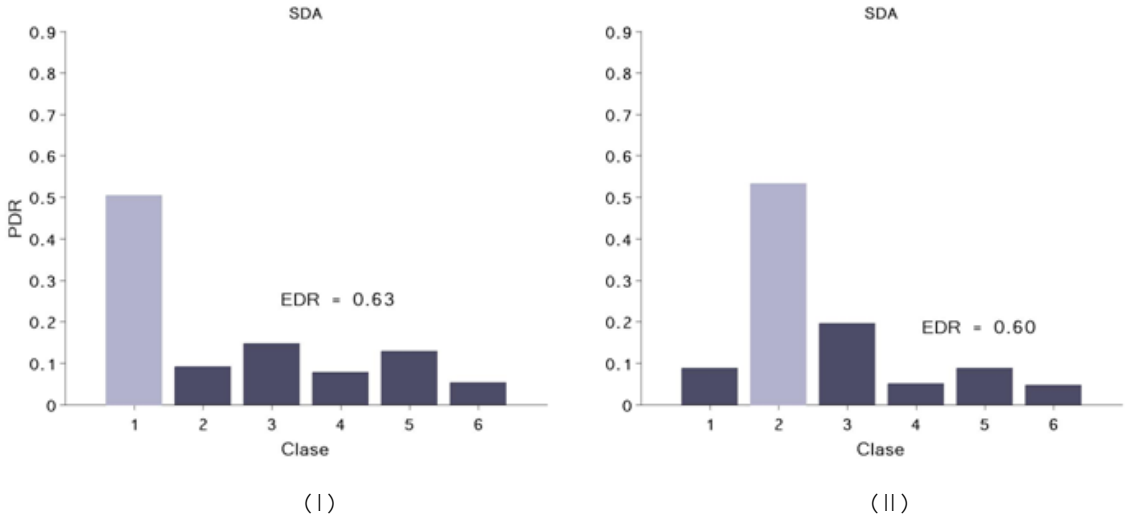


Figura 5.9. Cálculo de las PDRs y las EDRs usando el codificación cuaternaria. En cada gráfica la barra en gris claro representa la clase de las muestras comparadas con la librería. Una PDR grande significa una mayor cercanía a las muestras utilizadas respecto al resto de las clases. Cuanto menor es la EDR, más fácil es la discriminación porque una clase destaca sobre todas las demás por su cercanía a las muestras.



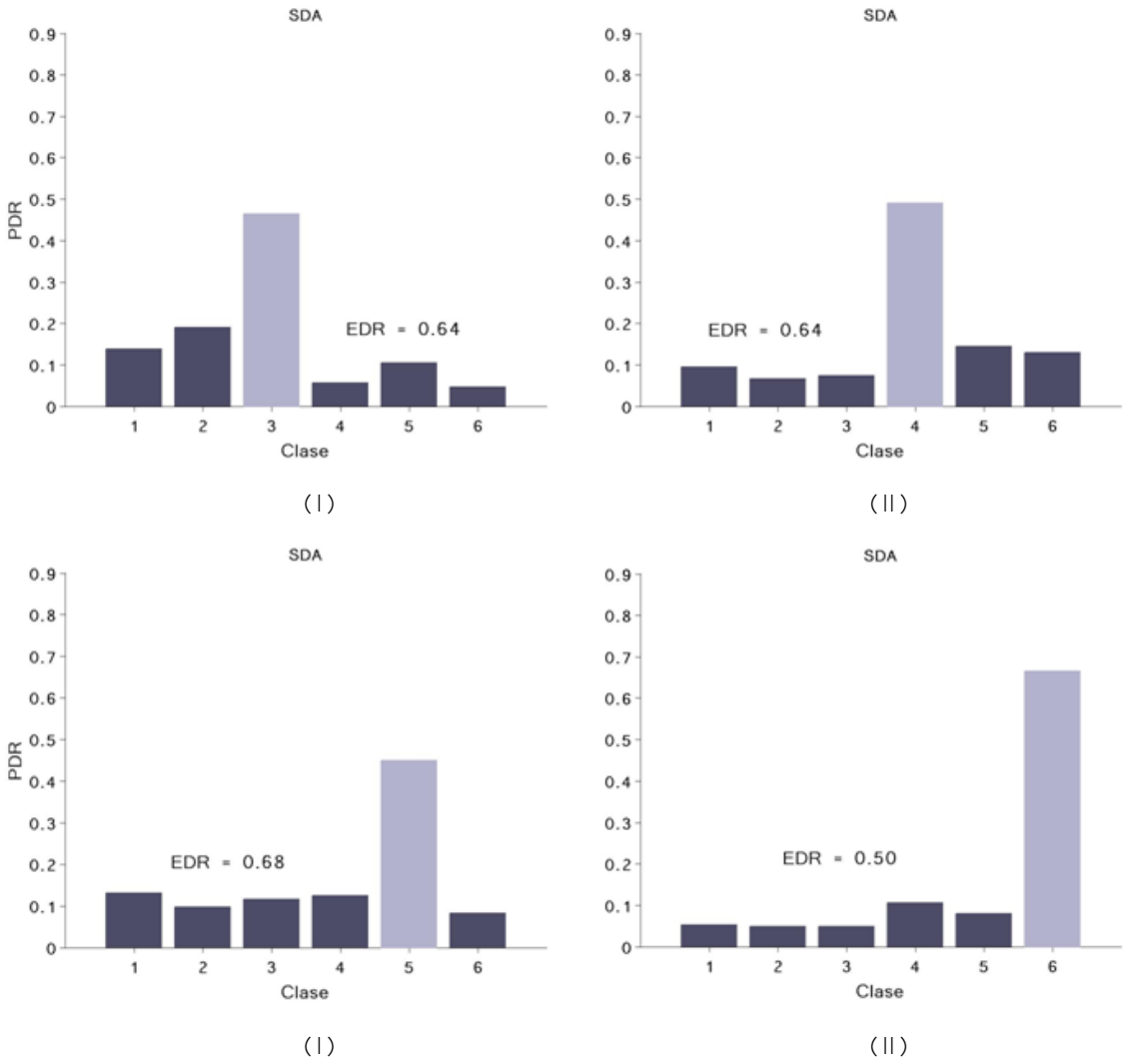


Figura 5.10. Cálculo de las PDRs y las EDRs usando medida espectral angular sobre la derivada (SDA). En cada gráfica la barra en gris claro representa la clase de las muestras comparadas con la librería. Una PDR grande significa una mayor cercanía a las muestras utilizadas respecto al resto de las clases. Cuanto menor es la EDR, más fácil es la discriminación porque una clase destaca sobre todas las demás por su cercanía a las muestras.

En base a los resultados parece que la medida de similitud que mejor consigue discriminar entre las especies es el coeficiente de correlación. Sin embargo hay que interpretar los datos con cautela. Por un lado, la correlación tanto para la especie correcta como para el resto devuelve distancias muy pequeñas, por tanto los índices relativos como PDR y EDR pueden dar lugar a valores elevados. Por otro, estos índices se han calculado después de obtener la distancia entre numerosas muestras para generalizar lo mejor posible, pero por ese mismo motivo no conocemos hasta qué punto puede haber muestras cuya menor distancia sea con un espectro de referencia distinto al correcto.

Dicho esto, las mejores conclusiones que se pueden extraer de este experimento son que tanto EDM, SAM como CC son aptas para ser utilizadas por los algoritmos de clasificación para medir la similitud o disimilitud entre los espectros de emisión de fluorescencia. Las condiciones para poder utilizar EDM son que los espectros estén normalizados y suavizados, mientras que usando la correlación no es tan vital. Como era de esperar las medidas de distancia basadas en codificaciones groseras de los datos no dan resultados suficientemente buenos como para que sean aplicados, pero sí sorprende la buena discriminación que ofrecen entre algunas de las especies. Por ejemplo *Thalasiossira weissflogii* y *Ostreococcus sp.* pueden ser fácilmente distinguidas del resto con una simple codificación binaria.

En general también se cumple que la especie que más causa confusión con Thwi es Pl, al igual que para Duna, mientras que para Pl la más cercana es esta última. Para Amin es Iso la más próxima y viceversa. En cuanto a Ost, esta se encuentra bastante separada del resto, pero las especies más cercanas serían Amin e Iso.

De la utilización de la primera derivada, en base a los pocos datos de los que se disponen, no parece que ofrezca mayor capacidad discriminativa que los datos originales, al menos al ser utilizados sin ningún tratamiento adicional.

5.5 Aprendizaje de distancias

Para disponer de una información visual de la separación de los datos que logran las medidas de distancia se utilizó un instrumento para proyectar los datos en dimensiones reducidas. Se trata de Multidimensional Scaling (Alpaydin, p.125), un método para disponer los datos en un espacio de por ejemplo dimensión dos, de manera que la distancia euclidiana entre los puntos preserve de la mejor manera posible la relación que entre ellos había en su dimensión original. Para ello se vale de una matriz en la que la única información de los datos es la distancia que posee cada muestra con respecto al resto, no importando la dimensión y el tipo de distancia utilizada.

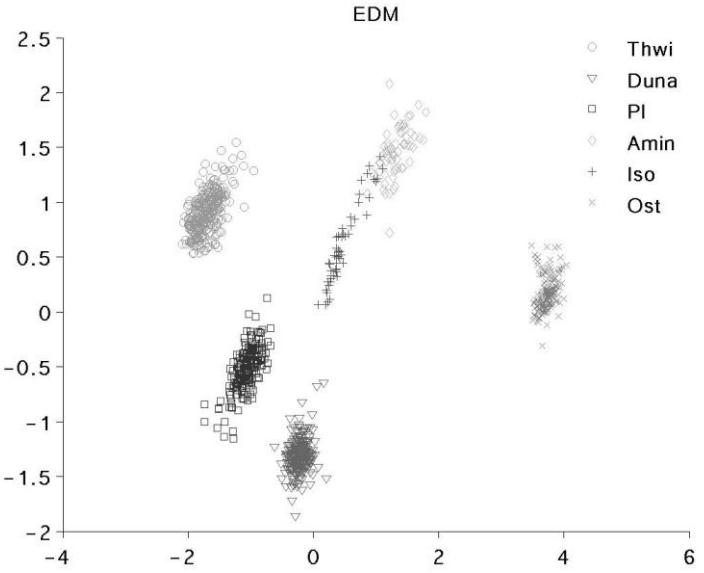


Figura 5.11. Representación en dos dimensiones de las muestras espectrales obtenida como resultado de aplicar Multidimensional Scaling sobre la matriz de distancias euclidianas mutuas entre todos los datos.

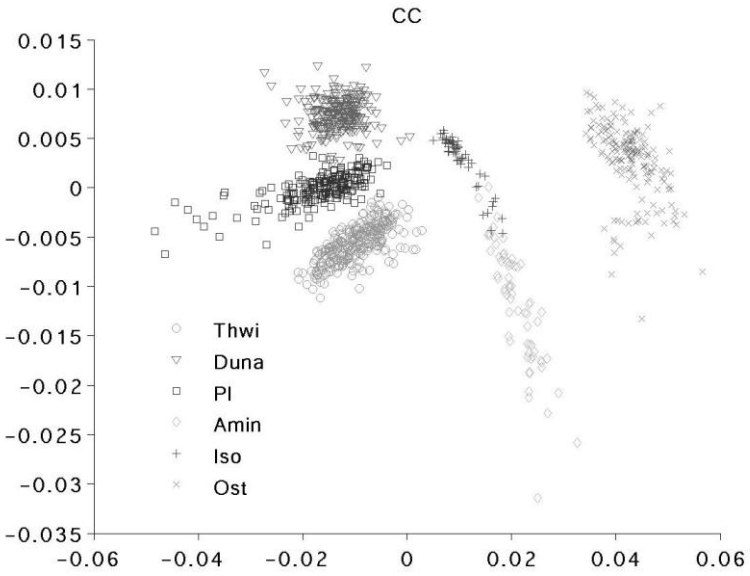


Figura 5.12. Representación en dos dimensiones de las muestras espectrales obtenida como resultado de aplicar Multidimensional Scaling sobre la matriz de coeficientes de correlación mutuas entre todos los datos.

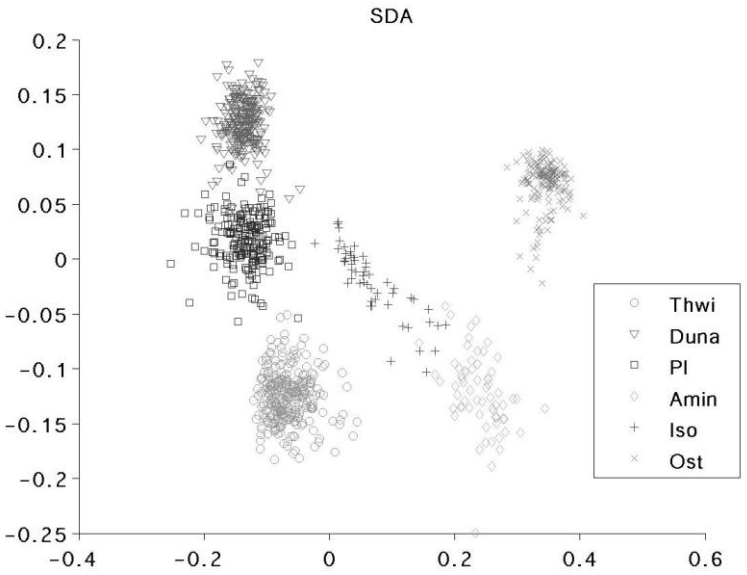


Figura 5.13. Representación en dos dimensiones de las muestras espectrales obtenida como resultado de aplicar Multidimensional Scaling sobre la matriz de distancias espectrales angulares mutuas entre todos la primera derivada de los datos.

Se confirman algunas de las conclusiones que se realizaron como la de cuáles son las especies más cercanas entre sí y además parece que la proyección de los datos a través de las distancias euclidianas es más espaciada. Los potenciales fallos en la clasificación son algunas muestras de Amin e Iso y otras de Thwi con Pl y de ésta última con Duna.

Con el fin de tratar de mejorar la separabilidad entre estas especies vamos a basarnos en las propiedades estadísticas que hemos intentado cuidar en el tema del pre-procesado. Tal como se hizo en el capítulo 3, volvemos a utilizar el test de Kolmogorov-Smirnov para verificar la normalidad de la distribución de las muestras en cada longitud de onda, pero esta vez aplicado a los datos bajo normalización wavelet y suavizado combinado.

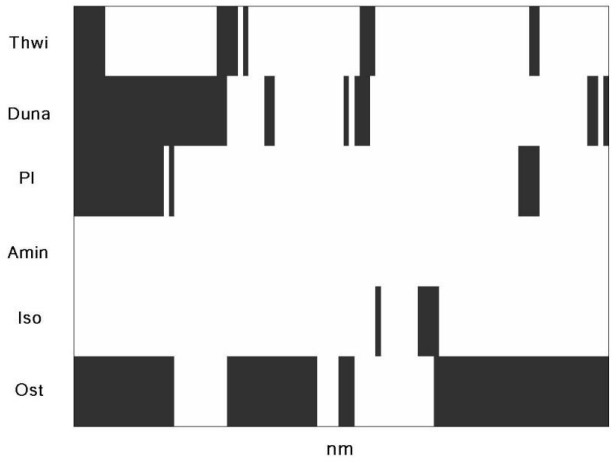


Figura 5.14. Resultado de aplicar el test de normalidad de Kolmogorov-Smirnov a las distintas especies en la banda 630-730 nm. En gris se muestran aquellas longitudes de onda para las que se ha rechazado la hipótesis de normalidad con un nivel de significación de 0.05.

Como las especies Amin e Iso son las que mejor han superado la prueba de normalidad, haremos el estudio con ellas. Veamos un ejemplo de las distribuciones de ambas especies a una longitud de onda concreta.

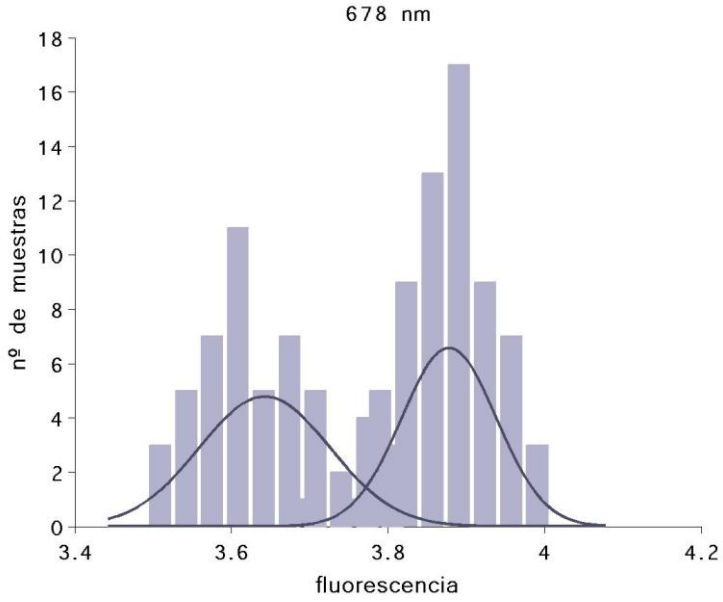


Figura 5.15. Histograma de las muestras de Amin e Iso a 678 nm. Las gaussianas extraídas de los parámetros de los histogramas aparecen superpuestas.

Dado que las distribuciones se pueden asemejar a la de una normal entonces a partir de la media y la varianza de los histogramas podemos crear una gaussiana para modelarlas. Si realizáramos una clasificación bayessiana para esta longitud de onda, para cada muestra seleccionaríamos la especie para la cual la probabilidad de pertenencia es mayor, es decir, aquella cuya gaussiana se encuentre por encima. También podemos calcular la probabilidad de errar en nuestra clasificación, que puede ocurrir si asignamos una muestra de Amin a Iso y viceversa.

$$\frac{1}{2}\left(\int_{-\infty}^{x_0} G_2\,dx + \int_{x_0}^{\infty} G_1\,dx\right) \tag{5.12}$$

Donde G es la ecuación de la gaussiana,

$$G = \frac{1}{\sqrt{2\pi}\sigma} e^{\left[-\frac{(x-\mu)}{2\sigma^2}\right]} \tag{5.13}$$

Si hacemos lo mismo para el resto de longitudes de onda contaremos con una estimación de las zonas en las que ambas curvas se encuentran separadas y aquellas otras en las que éstas se solapan.

Figura 5.16. Error de clasificación bayessiana de la banda 630-730 nm. Las líneas más oscuras indican mayor probabilidad de error.

Podemos aprovechar esta información para generar un vector de pesos para por ejemplo particularizar la distancia euclidiana a los datos. Las bandas con probabilidad de error cero le asignamos peso 1 mientras que las bandas con error mayor que 0.5 tendrían peso cero. Otras probabilidades recibirán pesos intermedios. Podemos incluso establecer un criterio más duro y quedarnos solo con las bandas cuya probabilidad de error no supere un valor establecido. Creando este vector de pesos W lo introduciríamos en la ecuación del cálculo de la distancia euclidiana (Xing et al. 2002):

$$\boldsymbol{WEM}(\mathrm{x},\mathrm{y}) = \sqrt{(\mathrm{x}-\mathrm{y})^{\mathrm{T}}\mathrm{W}(\mathrm{x}-\mathrm{y})} \tag{5.14}$$

Para probar su efectividad, calculamos la PDR entre las clases Amin e Iso:

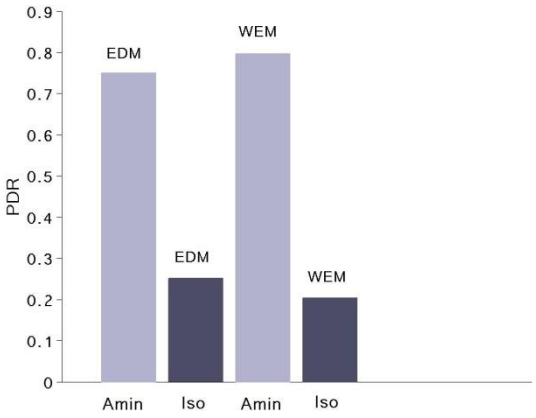


Figura 5.17. PDR entre las especies Amin e Iso. Para las dos primeras barras se utilizó la distancia euclidiana convencional (EDM) y para las dos últimas la distancia euclidiana ponderada (WEM).

Resumen

En el presente capítulo se han utilizado dos parámetros (PDR y EDR) para evaluar la discriminabilidad relativa que proporcionan algunos tipos de medida de distancia aplicables a espectros, para así conocer cual mantiene una mayor separabilidad entre los datos disponibles y cuáles son las especies más propensas a ser confundidas por un clasificador. Además se han proyectado algunas de las distancias en un espacio de dos dimensiones para confirmar las conclusiones extraídas con el anterior método.

Por último se desarrollado una forma de obtener un vector de pesos basado en la probabilidad de error en cada banda para adaptar la distancia euclidiana al caso específico del cálculo de distancia entre dos especies.

6. Clasificación de especies

6.1 Introducción

En los capítulos anteriores se han estudiado distintas formas de acondicionar los datos con la finalidad de facilitar el trabajo y mejorar el rendimiento del clasificador, algoritmo informático que tratará de identificar las especies a partir de los datos que han sido tratados. Es en la presente sección donde finalmente se averigua hasta qué punto la emisión de fluorescencia del fitoplancton se presta para ser utilizada en esta tarea.

Los experimentos desarrollados no pretenden ser un estudio exhaustivo de las posibilidades que ofrece el campo del *machine learning* o el aprendizaje de máquinas aplicado a nuestro problema, sino un instrumento a través del cual evaluar los métodos desarrollados. Sin embargo los métodos utilizados se han centrado en un tipo muy concreto, el de los mapas auto-organizados. *Self-organizing maps* (SOM) ya había sido probado anteriormente (Aymerich et al. 2009) para el mismo fin, pero las mejoras introducidas en el pre-procesado hacían conveniente realizar una comparación. Además los resultados que con él se obtienen también sirven de referencia para evaluar otro de los métodos utilizados, *Growing cell structures*, otro tipo de red auto-organizada pero que no tiene una estructura fija sino que crece para adaptarse a las necesidades que imponen los datos. Ambas técnicas serán introducidas más adelante.

Dado que los resultados del capítulo anterior relacionados con la capacidad discriminativa que ofrecen los distintos tipos de distancia dan a entender que los espectros tienen una separación suficientemente buena para la mayor parte de las especies empleadas, parece posible utilizar con buenos resultados un sencillo clasificador basado en distancia como k-vecinos, de forma que también sirva de referencia para los otros métodos.

Los principales resultados que se desean extraer son los índices de clasificación que se obtienen, qué método se comporta mejor utilizando datos hiperespectrales, y qué normalización es más conveniente.

En general SOM utiliza la distancia euclidiana dentro de su algoritmo para entrenar su red y clasificar los datos. Uno de los objetivos del proyecto era la de modificar su implementación para

alojar una distancia basada en ángulos para aprovechar la flexibilidad que ofrece con cierto tipo de datos.

6.2 Aprendizaje de máquinas

Simular el funcionamiento del cerebro humano es una de las grandes ambiciones del ser humano. Ésta sería la finalidad última del aprendizaje automático o de máquinas, un apasionante campo de estudio en el que se fusionan en mayor medida la informática, la estadística y el tratamiento de datos. El desarrollo del aprendizaje de máquinas ha sido posible en gran medida gracias a la mejora en la tecnología de la computación, procesadores más rápidos y memorias más grandes que permiten procesar el gran volumen de datos que los sistemas de aprendizaje demandan.

En general el aprendizaje automático consiste en un programa informático que es capaz de extraer información de conjunto de muestras para encontrar patrones en él que puedan ser de utilidad o bien para saber interpretar nuevas muestras que se le presenten en el futuro. Los dos grandes tipos de aprendizaje son el supervisado y el no-supervisado. En el aprendizaje supervisado es aquél en el que solo hay datos de entrada y el objetivo es buscar regularidades o patrones en los datos. Por ejemplo se pueden intentar agrupar aquellos que tengan características comunes

En nuestro caso lo que queremos es entrenar un modelo para que a partir de una serie de muestras hiperespectrales aprenda cuales son las características comunes entre las que pertenecen a una misma especie para permitir la identificación automática de nuevas muestras que el programa no haya visto anteriormente. Se trata de un aprendizaje supervisado, puesto que al conocer la etiqueta, en nuestro caso especie, a la que pertenece cada muestra, podemos supervisar el aprendizaje del modelo indicándole la salida que deseamos para cada entrada que se le presenta.

Para probar un algoritmo de aprendizaje necesitamos contar con un conjunto de entrenamiento para aprender un modelo, un conjunto de validación para seleccionar los parámetros del modelo más adecuado para los datos y un conjunto de test para estimar la probabilidad de error de clasificación. Para entrenar al clasificador es deseable contar con un gran número de datos para que el modelo aprendido sea capaz de generalizar mejor y no se especialice solo en los datos con los que fue creado. Para los conjuntos de validación y test también es deseable utilizar el mayor número de muestras posibles para que la selección de los hiperparámetros sea también la más adecuada en general y que la estimación del error sea

estadísticamente fiable. Sin embargo también es deseable generar varios clasificadores con diferentes datos para estudiar la repetibilidad.

En situaciones como la nuestra, en la que el número de datos es limitado debido al coste o la dificultad que supone extraerlas, hay que llegar a un compromiso en cuanto al número de datos que se utilizarán para cada conjunto. Lo que procede en estos casos es separar de antemano un cierto número de muestras para el conjunto de test. Con las restantes se realiza una validación cruzada que consiste en generar distintos conjuntos de entrenamiento y validación a partir del mismo set de datos. Una manera de realizarla es creando k grupos de los datos y utilizando k-1 para entrenar el primer clasificador y uno para la validación. Para el segundo clasificador se deja fuera del entrenamiento a otro grupo diferente y así sucesivamente hasta entrenar a k clasificadores. Se trata del *k-fold cross-validation*. Otro tipo de validación cruzada es el 5x2 *cross-validation*. Ésta consiste en dividir el conjunto de datos en dos y utilizar una mitad para el entrenamiento y la otra para validar. Para crear el segundo clasificador, la mitad que anteriormente sirvió para entrenar pasa a ser el conjunto de validación y viceversa. Para el tercer clasificador se realiza una nueva partición diferente del anterior puesto que se realiza de forma aleatoria. Este proceso se repite hasta que se efectúan 5 particiones diferentes, cada una de las cuales genera dos nuevos clasificadores. Si se efectuaran nuevas particiones los datos de éstas estarían ya demasiado solapados con las anteriores y la información calculada con ellos, como la probabilidad de error no aportan nueva información estadística (Alpaydin, p. 488).

En nuestro caso el método de separación de los datos en diferentes grupos de validación y test varía para cada especie debido a que para algunas de ellas hay un número de muestras muy dispar.

6.3 Técnicas de aprendizaje

6.3.1 K-vecinos

Uno de los clasificadores utilizados es uno de los más simples. Se trata de k-vecinos (Alpaydin, p.172) más cercanos, siendo k un parámetro de diseño. Ni siquiera requiere entrenamiento, solo un conjunto de muestras de las distintas clases sobre las que calcula la distancia con la muestra que se desea clasificar. La clase más representada dentro de las k muestras con las que tiene menor distancia, será la clase asignada. Tiene una fuerte dependencia con el conjunto de datos de referencia y el tipo de distancia utilizada.

6.3.2 Self-Organizing Maps

La red neuronal auto-organizada (Self-Organizing Maps - SOM) (Kohonen), es un tipo de red neuronal artificial con aprendizaje del tipo no supervisado que crea una representación (o mapa) en dimensiones bajas, usualmente dimensión dos, del espacio de los datos de entrada. Las dos propiedades que caracterizan SOM y la diferencian de otras redes neuronales es por un lado su carácter competitivo y el hecho de que el aprendizaje esté reservado especialmente para las neuronas que ya de por sí exhiben un mayor parecido con la entrada o estímulo.

El mapa está formado por neuronas (o nodos) a las que les son asignadas un vector de la misma dimensión que los datos de entrada, además de una posición dentro del mapa definida por las neuronas vecinas con las que conecta. El entrenamiento es un proceso competitivo en el sentido de que las neuronas más parecidas a la entrada y sus vecinos son las que modificarán sus valores para acercarse más a ella.

Como resultado los nodos terminan con una estructura ordenada, de forma que aquellas próximas entre sí dentro del mapa tienen vectores de pesos similares. De la misma forma, muestras similares dentro del espacio de los datos de entrada son proyectadas cerca y las disimilares separadas, preservando así las propiedades topológicas del espacio de entrada. Este principio de la preservación de la topología es un principio importante dentro de los sistemas de procesados de señales biológicas.

SOM puede interpretarse como una forma de condensar los datos ya que crea un conjunto de vectores prototipo, cuyo número está definido por el tamaño de la red, que representan a los datos de entrada. Es esta característica lo que hace de SOM un algoritmo adecuado para realizar clasificaciones rápidas. Otras formas de entender el mapa que genera SOM es como un grafo de similaridad o un diagrama de agrupamiento.

El aprendizaje de la red es un proceso iterativo. En cada paso del entrenamiento una muestra del conjunto de datos es seleccionada y las distancias, habitualmente la euclidiana, entre ésta y todos los nodos del mapa son calculadas. La neurona ganadora (*Best Matching Unit* - BMU) es aquella que posee la menor distancia, y la adaptación o aprendizaje de ésta y el resto de nodos está gobernada por la siguiente ecuación:

$$m_i(n+1) = m_i(n) + \alpha(n)h_{ci}(n)[x(n) - m_i(n)] \qquad (6.1)$$

- i es el índice del nodo.

- n es el paso de aprendizaje actual.
- h es el kernel de vecindad definido alrededor del nodo ganador c.
- x es la muestra de entrada.
- α es el factor de aprendizaje.

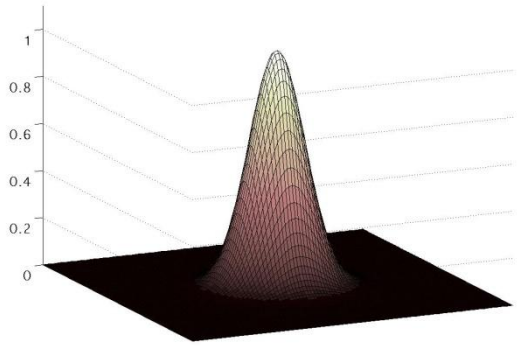


Figura 6.1. Kernel gaussiano, un tipo de función utilizada para definir el aprendizaje de la vecindad del nodo ganador.

El kernel de vecindad es la función que define cuánto aprende una neurona respecto a las que tiene a su alrededor. Es una función decreciente centrada en el nodo ganador y que decae con la distancia respecto a él. Habitualmente se utiliza la función gaussiana. El parámetro α toma valores entre 1 y cero, decreciendo a medida que avanza el aprendizaje para pasar de un aprendizaje global para que todo el mapa adquiera la estructura de los datos a una fase de aprendizaje local para que cada zona se especialice en una determinada forma o patrón de los datos. Por tanto, los dos parámetros que fijan el ámbito dentro del mapa sobre el cual cada muestra de entrada va a tener influencia son α y el ancho del kernel utilizado (desviación típica en el caso de utilizar una gaussiana).

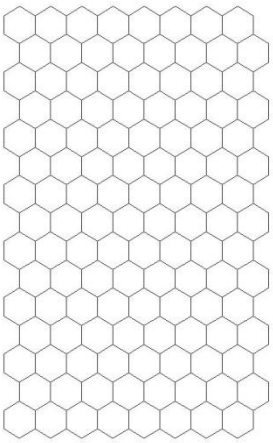


Figura 6.2. Ejemplo de mapa o grid donde cada hexágono alojaría una neurona. Por tanto cada uno de ellos tendría 6 nodos vecinos, excepto los que se encuentran en los bordes.

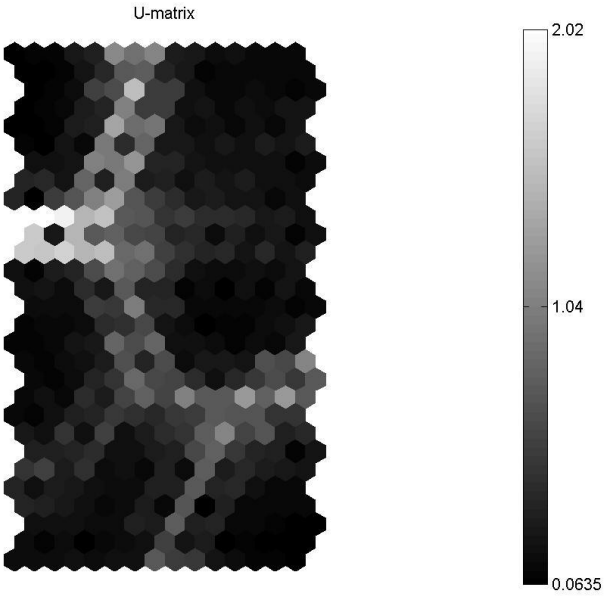


Figura 6.3. Ejemplo de matriz U obtenida al entrenar el mapa de SOM con seis especies de fitoplancton. Cuánto más oscuro esté representado un hexágono menor distancia hay entre los nodos en esa zona. Las zonas claras representan fronteras entre los agrupamientos de las clases.

Una vez que se ha completado el entrenamiento se pueden obtener visualizaciones del mapa para obtener información relativa a los agrupamientos de los datos y la separabilidad entre ellos. Por ejemplo se puede visualizar la matriz U en cuya representación se muestran las distancias entre cada nodo y sus vecinos mediante un código de colores. En ella se pueden identificar fácilmente fronteras entre agrupamientos (clusters) de los datos (figura 6.3).

También se puede calcular la *hit* matrix. En esta matriz se representan los nodos ganadores de los datos que se utilizaron para entrenar la red. A partir de ella también se puede obtener una visualización del mapa en la que los nodos que han sido ganadores para una determinada especie son etiquetados con su nombre (figura 6.4). Comparando las dos matrices podemos asociar las agrupaciones visibles en la matriz U con la especie a la que pertenece. Como ya se destacó en el capítulo anterior, nuevamente se observa que las fronteras más difusas son las que hay entre Amin e Iso y entre Duna y Pl.

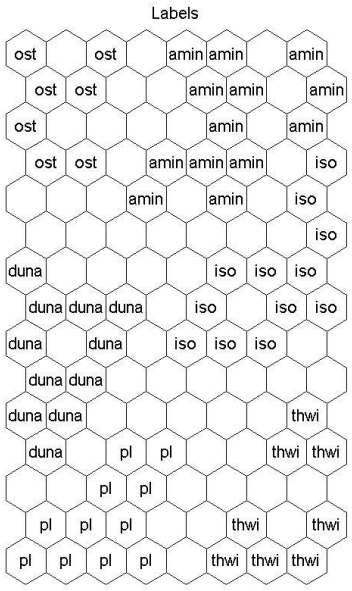


Figura 6.4. Ejemplo de *hit matrix* para el mismo mapa que el de la figura 6.3.

Para hacer de la red entrenada de SOM un clasificador supervisado se calcula una matriz de pertenencia. De forma similar a como se obtuvo la *hit matrix*, se obtienen los nodos ganadores de las muestras de entrenamiento, pero a diferencia del caso anterior a los nodos vecinos también se les asigna un grado de pertenencia, menor cuánto más alejado esté del ganador. De esta forma

cada muestra activa una cierta zona del mapa y la similitud con cada nodo se define de forma difusa y no discreta como en la matriz h. Así cada nodo se asignará a la especie cuyo grado de pertenencia sea mayor. Para clasificar basta calcular la distancia de la muestra de test con todos los nodos y etiquetarla con la especie asignada al nodo ganador.

El algoritmo de entrenamiento y clasificación con SOM se resume en los siguientes pasos:

- Inicializar los nodos del mapa.
- Para cada muestra obtener el nodo ganador y adaptar éste y su vecindad una cantidad proporcional a la diferencia con el dato de entrada, el valor que toma el kernel de vecindad y el factor de aprendizaje.
- Repetir el paso anterior las veces que sean necesarias hasta completar el entrenamiento.
- Obtener la matriz de pertenencia que generan las muestras de entrenamiento etiquetadas sobre los nodos de la red.
- Para cada muestra de test calcular el nodo ganador y asignarle la clase que éste tiene asignado.

6.3.3 SOM angular

Como muchos algoritmos de agrupamiento y de clasificación, Self Organizing Maps (SOM) depende fuertemente de la medida utilizada para obtener la distancia entre las muestras y los nodos que forman parte de la red, por lo que resulta crucial hacer uso de aquella que mejor se adapte a las características de los datos. La medida de similitud habitualmente usada en SOM es la euclidiana. Como ya se ha indicado, uno de los inconvenientes de esta distancia es su dependencia con la escala de los datos. Por ejemplo en el caso de los espectros de fluorescencia, dos medidas sobre el mismo cultivo efectuada con una configuración diferentes de los parámetros o con diferentes instrumentos dará lugar a escalados distintos. Una manera de eliminar su influencia es utilizar medidas de distancia angulares, por definición independientes de la escala. Para ello se planteó la modificación del algoritmo de SOM para que albergue la posibilidad de utilizar este tipo de distancia. En concreto se utilizará el Spectral Angle Measure (SAM), ya presentado en el capítulo 5.

Las fases del algoritmo en las que es necesario medir la distancia entre vectores de datos son, en primer lugar, la etapa de búsqueda durante del BMU durante la fase de entrenamiento. En segundo lugar, durante la fase de clasificación, también se hace el cómputo de la distancia entre

las muestras de entrenamiento y el mapa ya entrenado para obtener la matriz de pertenencia de cada nodo a cada clase y finalmente para encontrar el nodo ganador de las muestras de test.

Además, se ha de tener en cuenta que durante el aprendizaje, cada vez que se obtiene el BMU de una muestra, éste y sus vecinos se adaptan para parecerse más a los datos:

$$m_i(n+1) = m_i(n) + \alpha(n)h_{ci}(n)[x(n) - m_i(n)] \qquad (6.2)$$

Cuando la medida de distancia utilizada para esta búsqueda es la euclidiana, entonces resulta lógico que esta adaptación sea proporcional a la diferencia entre el vector de pesos del nodo y los de la muestra. Sin embargo, es necesario modificar la aproximación progresiva de los nodos a los datos de forma que sea coherente con la nueva medida de distancia que se está aplicando.

El cambio de medida euclidiana a angular para la búsqueda del BMU es inmediato, pues basta aplicar de forma directa la ecuación de *Spectral Angle Measure* (SAM). Sin embargo para poder realizar la adaptación de los nodos se pueden seguir dos estrategias distintas:

- Una vez calculado el ángulo en coordenadas cartesianas, transformar los vectores de la muestra y de todos los nodos a coordenadas hiperesféricas. A continuación calcular la adaptación de los vectores de pesos de la red proporcional a la diferencia entre las componentes angulares de la muestra y los nodos. Finalmente transformar de nuevo a coordenadas cartesianas.
- Antes de comenzar el entrenamiento, transformar los vectores de los nodos y de los datos de entrenamiento a coordenadas hiperesféricas. Para la búsqueda del BMU, en lugar de aplicar la ecuación de SAM, se hace uso de las ecuaciones que calculan el ángulo formado por dos vectores en coordenadas hiperesféricas. Se adaptan los nodos procediendo de igual forma al caso anterior y por último, una vez finalizado el proceso de aprendizaje, se invierte la transformación inicial.

Las coordenadas hiperesféricas son la generalización de las coordenadas esféricas en tres dimensiones a dimensión n, y por tanto cuentan con 1 coordenada radial r, llamada hiperradio, y $n-1$ ángulos hiperesféricos ϕ. Las ecuaciones para pasar de un sistema de coordenadas a otro son las que siguen:

$$r = \sqrt{x_n^2 + x_{n-1}^2 + \cdots + x_2^2 + x_1^2} \qquad (6.3)$$

$$\phi_1 = arccot\frac{x_1}{\sqrt{x_n^2 + x_{n-1}^2 + \cdots + x_2^2}} \tag{6.4}$$

$$\phi_2 = arccot\frac{x_2}{\sqrt{x_n^2 + x_{n-1}^2 + \cdots + x_3^2}} \tag{6.5}$$

$$\vdots$$

$$\phi_{n-2} = arccot\frac{x_{n-2}}{\sqrt{x_n^2 + x_{n-1}^2}} \tag{6.6}$$

$$2arccot\frac{\sqrt{x_n^2 + x_{n-1}^2} + x_{n-1}}{x_n} \tag{6.7}$$

Y la transformada inversa:

$$x_1 = r \cdot \cos(\phi_1) \tag{6.8}$$

$$x_2 = r \cdot \sin(\phi_1)\cos(\phi_2) \tag{6.9}$$

$$x_3 = r \cdot \sin(\phi_1)\sin(\phi_2)\cos(\phi_3) \tag{6.10}$$

$$\vdots$$

$$x_{n-1} = r \cdot \sin(\phi_1) \ldots \sin(\phi_{n-2})\cos(\phi_{n-1}) \tag{6.11}$$

$$x_n = r \cdot \sin(\phi_1) \ldots \sin(\phi_{n-2})\sin(\phi_{n-1}) \tag{6.12}$$

Para el cálculo del ángulo formado por dos vectores $\vec{u}, \vec{v}$ en coordenadas hiperesféricas se hace uso de una generalización de la fórmula de Harvesine. A partir de su versión en cartesianas:

$$\alpha = arccos\frac{N}{D} \tag{6.13}$$

$$N = u_1v_1 + u_2v_2 + \cdots + u_nv_n \tag{6.14}$$

$$D^2 = (u_1^2 + u_2^2 + \cdots + u_n^2)(v_1^2 + v_2^2 + \cdots + v_n^2) \tag{6.15}$$

Se obtiene la versión para hiperesféricas:

$$N = |\vec{u}| \cdot sin(u_{\phi_1}) \cdots \sin(u_{\phi_{n-1}}) \cdot |\vec{v}| \cdot \sin(u_{\phi_1}) \cdots \sin(u_{\phi_{n-1}}) + \tag{6.16}$$

$$+|\vec{u}| \cdot sin(u_{\phi_1}) \cdots \sin(u_{\phi_{n-2}}) \cdot \cos(u_{\phi_{n-1}}) \cdot |\vec{v}| \cdot sin(u_{\phi_1}) \cdots \sin(u_{\phi_{n-2}}) \cdot \cos(u_{\phi_{n-1}}) + \cdots$$

$$\cdots + |\vec{u}| \cdot cos(u_{\phi_1}) + |\vec{v}| \cdot cos(u_{\phi_1})$$

$$D = \left[\left(|\vec{u}| \cdot sin(u_{\phi_1}) \cdots \sin\left(u_{\phi_{n-1}}\right) \right)^2 + \left(|\vec{u}| \cdot sin(u_{\phi_1}) \cdots \sin\left(u_{\phi_{n-2}}\right) \cdot \cos\left(u_{\phi_{n-1}}\right) \right)^2 + \cdots \right. \quad (6.17)$$

$$\left. \cdots + \left(|\vec{u}| \cdot cos(u_{\phi_1}) \right)^2 \right] \cdot \left[\left(|\vec{v}| \cdot \sin\left(u_{\phi_1}\right) \cdots \sin\left(u_{\phi_{n-1}}\right) \right)^2 + \right.$$

$$\left. + \left(|\vec{v}| \cdot sin(u_{\phi_1}) \cdots \sin\left(u_{\phi_{n-2}}\right) \cdot \cos\left(u_{\phi_{n-1}}\right) \right)^2 + \cdots + \left(|\vec{v}| \cdot cos(u_{\phi_1}) \right)^2 \right]$$

6.3.4 Growing Cell Structures

Growing cell structures (GCS) (Fritzke 1994) es, al igual que SOM, una red neuronal auto-organizada. Usado como método no supervisado sus principales aplicaciones son la visualización de datos, el agrupamiento y la cuantificación vectorial. La principal ventaja sobre otros métodos existentes, como puede ser SOM, es su habilidad para obtener automáticamente una estructura y un tamaño de red adecuado para cada tipo de datos. Esto se consigue a través de un proceso de crecimiento controlado que también puede incluir la eliminación ocasional de unidades superfluas. El resultado son redes de pequeño tamaño, que generalizan muy bien y cuya principal aplicación es la clasificación automática.

Se ha constatado que la estructura y el tamaño predefinidos de los mapas de SOM suponen una limitación para las proyecciones resultantes (Fritzke 1993). Por otro lado, en la mayoría de los casos no se dispone de información a priori que permita elegir de antemano un tamaño y una forma idóneos. Una solución para este problema es determinar la forma así como el tamaño durante la simulación dentro de un modelo progresivo. Éste es el principio que lleva a cabo GCS de forma que se obtengan estructuras flexibles y dependientes del problema.

El proceso de entrenamiento de GCS es bastante similar al de SOM. La principal diferencia es que cada cierto número de pasos se inserta un nuevo nodo allí donde el algoritmo lo considera más necesario. Cada nodo tiene asociado una variable de error que aumenta una cantidad proporcional a la diferencia entre éste y el vector de la muestra de entrada cuando ha resultado ser el nodo ganador. Después de un cierto número de etapas de adaptación preestablecido, se busca el nodo que en el proceso ha acumulado más error. Entre éste y el vecino con el que tiene mayor distancia se inserta uno nuevo cuyo vector se inicializa con la media de los vectores de los nodos entre los que se creó, además de heredar los vecinos comunes a ambos.

Otra diferencia respecto a SOM es la forma en la que realiza la adaptación de los vectores nodales. GCS solo adapta al nodo ganador y a los vecinos con los que está conectado. Para el ganador existe un parámetro de aprendizaje diferente que para los vecinos:

$$m_c(n+1) = m_c(n) + eb(x - m_c(n)) \qquad (6.18)$$

$$m_i(n+1) = m_i(n) + en(x - m_i(n)) \quad (para\ i\ \in N_r) \qquad (6.19)$$

- c es el nodo ganador.
- x es el vector de entrada.
- eb es el factor de aprendizaje del nodo ganador.
- en es el factor de aprendizaje de sus vecinos.
- N_r es el conjunto de nodos que son vecinos de c.

Como se ve, GCS no tiene un decrecimiento del ámbito del aprendizaje como tenía SOM. La especialización de las neuronas se logra mediante la introducción de nuevos componentes de la red.

La fase de validación se realiza justo antes de cada introducción de un nuevo nodo para evaluar el desempeño de la red entrenada hasta entonces y poder utilizar, si se desea, el error de clasificación como criterio de parada. De forma similar a SOM, las muestras de entrenamiento también generan una probabilidad de que cada nodo pertenezca a una clase en particular. En la implementación utilizada, la clase asignada a la muestra de test no es directamente la asociada con el nodo ganador, sino que generará una cierta activación en cada nodo en función de la distancia que tenga con ellos modelado por una gaussiana, que ponderado por las probabilidades de pertenencia dan lugar a la etiqueta más probable.

El algoritmo de clasificación de GCS se resume en las siguientes líneas:

- Inicializar la red con k nodos.
- Para cada muestra de entrada calcular el nodo ganador y adaptar éste y sus vecinos directos una cantidad proporcional a la diferencia con la muestra y al factor de aprendizaje asociado con cada uno.
- Incrementar la variable de error del nodo ganador.
- Una vez se ha llevado a cabo el número de adaptaciones preestablecida, insertar un nuevo nodo entre aquel que haya acumulado mayor error y su vecino más distante.
- Calcular las probabilidades de pertenencia de los nodos a cada especie.

- Para cada muestra de test calcular la activación que produce sobre los nodos de la red y ponderarla con las probabilidades de pertenencia para hallar la etiqueta más verosímil.
- Repetir hasta que se hayan insertado el número de nodos deseado o bien se haya cumplido algún otro criterio de parada.

6.4 Metodología y resultados

Para facilitar la comparación entre las distintas técnicas se ha de utilizar algún índice que aporte información sobre la probabilidad que tiene el clasificador de equivocarse. Los instrumentos utilizados para tal fin han sido la matriz de confusión y el índice Kappa. La matriz de confusión permite visualizar tanto el número de muestras que han sido correctamente clasificadas como las que no, y para estas últimas permite conocer con qué otra clase se han confundido. El índice Kappa es una medida de error global en la clasificación cuyo cálculo se efectúa a partir de los elementos de la matriz de confusión de la siguiente forma:

$$Kappa = \frac{n\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i+}n_{+i}}{n^2 - \sum_{i=1}^{k} n_{i+}n_{+i}} \qquad (5.8)$$

- n es el número total de muestras utilizadas para clasificar.
- k es el número de especies.
- n_{ii} es el número de elementos clasificados correctamente para la especie i, es decir los distintos componentes de la diagonal principal.
- n_{i+}es la suma de los elementos de la fila i, es decir el número total de muestras de la especie i utilizadas para clasificar.
- n_{+i} es la suma de los elementos de la columna i, es decir el número total de muestras que el clasificador ha asignado la etiqueta de la especie i.

Como ya se indicó, la manera de realizar experimentos estadísticamente fiables con un número limitado de datos es realizando una validación cruzada. En primer lugar resumimos los datos con los que contamos después de haber impuesto la condición de que tengan una unidad mínima de 1 unidad de fluorescencia:

Especie	Clase	Abreviación	Nº de muestras (1ª toma)	Nº de muestras (2ª toma)

Thalassiosira weissflogii	Bacillariophyceae	Thwi	300	26
Dunaliella primolecta	Chlorophyceae	Duna	249	26
Pleurochrysis elongata	Primnesiophyceae	Pl	201	25
Alexandrium minutum	Dinophyceae	Amin	65	14
Isochrysis Galbana	Primnesiophyceae	Iso	0	52
Ostreococcus sp.	Prasinophyceae	Ost	180	0

Tabla 6.1. Resumen de las muestras disponibles para formar los conjuntos de entrenamiento, validación y test.

Como se ve, las muestras disponibles de cada especie son muy dispares. En general hay demasiado pocas para hacer un *k-fold cross-validation*. Si usáramos directamente la validación cruzada 5x2, en cada partición habría especies sobre-representadas. Por ejemplo Thwi siempre tendría una presencia importante mientras que Iso tendría pocas muestras. No interesa que haya ninguna especie con más muestras que las otras durante la fase de entrenamiento porque, por ejemplo en el caso de SOM, abarcaría un mayor espacio de representación en el mapa, pudiendo perjudicar a la generalización del resto.

Como solución se optó por fijar un número fijo de muestras de cada especie para el entrenamiento, y por tanto para la validación. Por un lado las muestras del dataset correspondiente a la segunda toma de muestras se reservan para el test final. Como Ost solo se adquirió en la primera, se separan 20 muestras de las que serán utilizadas para entrenar, mientras que las mismas muestras de Iso que se utilizan para entrenar y validar también se utilizarán para el test.

La especie con menor número de muestras es Iso, por lo que será ésta la que fije el tamaño de las particiones. Para cada especie se formarán cinco grupos de 52 muestras, cada uno de los cuales servirá para entrenar dos clasificadores. Para formar los grupos de una especie se divide el número total de muestras que de él se dispone por los grupos a formar. El resultado redondeado a la baja es el número de muestras que no se repetirán en ninguno de ellos. Thwi es la única especie que puede formar 5 agrupaciones sin repetir ninguna muestra. Para el resto, a cada grupo

se le añade el número de muestras necesarias para completarlo tomadas aleatoriamente de las restantes.

La implementación utilizada de los diferentes métodos han sido la SOM toolbox para Matlab 5 para SOM, la toolbox gcssoft para GCS y prtools para la implementación de k-vecinos.

En primer lugar vamos a estudiar para GCS la dependencia del error de clasificación con el número de nodos. Se parte de una red de tres nodos inicializada con tres muestras aleatorias del conjunto de muestras. En la figura 6.5 se muestra la media del índice kappa en función del número de nodos como resultado de hacer crecer la red hasta que tuviera 50.

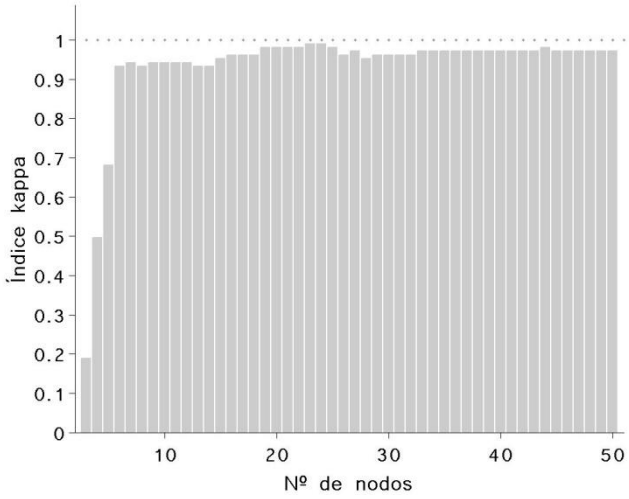


Figura 6.5. Evolución del índice kappa en función del número de nodos de una red de GCS usando datos normalizados mediante wavelet.

Vemos que la red alcanza una etapa de estabilidad después de unas primeras etapas en la que la red todavía no está adaptada a los datos. Veamos qué ocurre si inicializamos la red con seis nodos cuyos vectores toman los valores del espectro medio de cada clase (Figura 6.6).

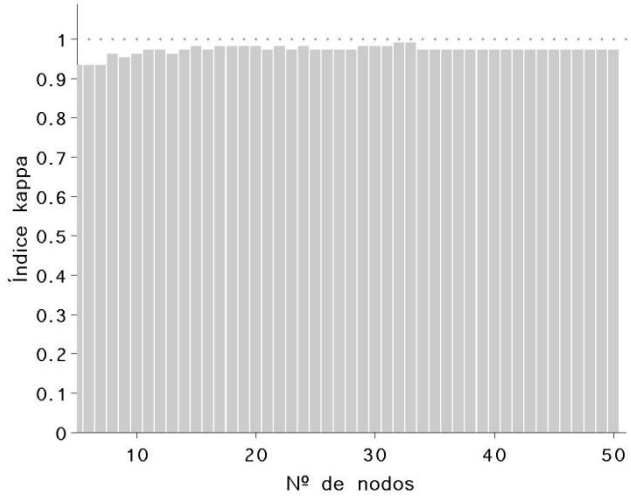


Tabla 6.6. Evolución del índice kappa en función del número de nodos de una red de GCS usando datos normalizados mediante wavelet.

La red comienza con una buena aproximación de cada especie y por tanto comienza con buenos resultados. Como la precisión en la clasificación es bastante estable con el tamaño de la red, para la comparación con SOM se va a utilizar un tamaño de 35 nodos, cantidad que está por encima de la fase inicial más inestable y además mantiene un número reducido de nodos. En teoría lo que se haría es, con ayuda del conjunto de validación, entrenarlo hasta que obtengamos la máxima precisión posible en la clasificación. Sin embargo, para que la comparación con SOM sea más justa fijamos un tamaño, dado que en teoría podríamos realizar la misma búsqueda del tamaño de mapa ideal para SOM, aunque se efectuara de una forma menos eficiente que con GCS. De todas formas también se incluirán los resultados para el mejor tamaño de red con GCS.

Las tablas de la 6.2 a la 6.9 muestran los resultados más relevantes.

	Max-Min	Crecimiento	Media y Varianza	Wavelet
2NN	0.987	0.984	0.987	**0.999**
SOM	0.935	0.904	0.953	**0.958**
GCS	0.965	0.974	0.969	**0.990**

Tabla 6.2. Índices Kappa obtenidos al clasificar usando distintas normalizaciones: Máximo y mínimo, modelado del crecimiento, media y varianza y wavelet.

	Orden 1	Orden 2	Orden 3	Orden 4	Orden 5
2NN	**0.988**	0.983	0.981	0.976	0.972
SOM	0.95	0.951	**0.952**	0.949	0.949
GCS	0.973	**0.978**	0.962	0.963	0.963

Tabla 6.3. Índices Kappa obtenidos al clasificar usando los espectros derivados para orden 1 hasta 5 sobre los datos normalizados con wavelet y suavizado con wavelet + SG.

	Selección de Variables	Extracción de Variables
2NN	**1**	0.982
SOM	0.970	**0.973**
GCS	**0.985**	0.980

Tabla 6.4. Índices Kappa obtenidos al clasificar usando las variables seleccionadas por el algoritmo genético (selección de variables) o extraídas por PCA (extracción de variables).

Max-Min	Crecimiento	Media y Varianza	Wavelet	Derivada Orden 1	Derivada Orden 2	Selección	Extracción
0.973	0.985	0.977	**0.994**	0.983	0.982	0.990	0.991

Tabla 6.5. Índices Kappa obtenidos al clasificar con GCS escogiendo para cada pareja de conjuntos de entrenamiento y validación, el tamaño de la red que mayor precisión de clasificación tenía.

2NN	Thwi	Duna	PI	Amin	Iso	Ost
Thwi	26	0	0	0	0	0
Duna	0	26	0	0	0	0
PI	0	0	26	0	0	0
Amin	0	0	0	25.2	**0.8**	0
Iso	0	0	0	0	26	0
Ost	0	0	0	0	0	26

SOM	Thwi	Duna	PI	Amin	Iso	Ost
Thwi	26	0	0	0	0	0
Duna	0	26	0	0	0	0
PI	0	0	26	0	0	0
Amin	0	0	0	26	0	0
Iso	0	0	0	**5.4**	20.6	0
Ost	0	0	0	0	0	26

GCS	Thwi	Duna	PI	Amin	Iso	Ost
Thwi	26	0	0	0	0	0
Duna	0	25.9	0	0	0.1	0
PI	0	0	26	0	0	0
Amin	0	0	0	25.3	**0.7**	0
Iso	0	0	0	0.2	25.8	0

Ost	0	0	0	0	0	26

Tabla 6.6. Matrices de confusión media de 2NN (a), SOM (b), y GCS (c) usando normalización wavelet.

2NN	Thwi	Duna	Pl	Amin	Iso	Ost
Thwi	26	0	0	0	0	0
Duna	0	25.9	0.1	0	0	0
Pl	0.3	0.3	25.4	0	0	0
Amin	0	0	0	25.3	**0,7**	0
Iso	0	0	0	0,2	25.8	0
Ost	0	0	0	0	0	26

2NN	Thwi	Duna	Pl	Amin	Iso	Ost
Thwi	26	0	0	0	0	0
Duna	0	25.7	0.2	0	0.1	0
Pl	0.4	0.5	25.1	0	0	0
Amin	0	0	0	25.2	**0.8**	0
Iso	0	0	0	0.2	25.8	0
Ost	0	0	0	0	0	26

Tabla 6.7. Matrices de confusión media de 2NN usando la derivada de orden 1 (a) y 2 (b).

SOM	Thwi	Duna	Pl	Amin	Iso	Ost
Thwi	26	0	0	0	0	0
Duna	0	26	0	0	0	0

PI	0.1	0	25.9	0	0	0
Amin	0	0	0	25.8	0.2	0
Iso	0	0	0	**3.2**	22.8	0
Ost	0	0	0	0	0	26

GCS	Thwi	Duna	PI	Amin	Iso	Ost
Thwi	26	0	0	0	0	0
Duna	0	25.7	0,1	0	0,2	0
PI	0	0	26	0	0	0
Amin	0	0	0	25.3	**0,7**	0
Iso	0	0	0	0,5	25.5	0
Ost	0	0	0	0	0	26

Tabla 6.8. Matrices de confusión media de GCS y SOM usando la selección de bandas del algoritmo genético (5 bandas).

2NN	Thwi	Duna	PI	Amin	Iso	Ost
Thwi	26	0	0	0	0	0
Duna	0	26	0	0	0	0
PI	0	0	26	0	0	0
Amin	0	0	0	25.2	0.8	0
Iso	0	0	0	1.5	24.5	0
Ost	0	0	0	0	0	26

Tabla 6.9. Matriz de confusión media de 2NN usando las variables extraídas por PCA (3 variables).

Lo primero que se observa en general es que los índices de precisión en la clasificación son muy altos. Este hecho era de esperar teniendo en cuenta los resultados de discriminación entre especies obtenidos en el capítulo 4. No sorprende tampoco que los mejores resultados se obtengan un clasificador tan sencillo como k-vecinos, al estar basado en distancias y tomando como referencia las muestras disponibles, similar a como se procedió en el citado estudio. Los otros dos clasificadores también usan de alguna manera espectros de referencia, pero al hacer una condensación de los datos, el resultado depende de lo bien que se haya realizado. Dada la igualdad entre los índices Kappa, la comparación de los resultados es difícil puesto que la diferencia entre uno u otro es muy pequeña. De todas formas sí hay ciertos patrones que merecen la atención del lector.

En la tabla 6.2 se observa que para los tres clasificadores la normalización para la cual se obtienen mejores resultados es la propuesta basada en la descomposición con wavelet. A pesar de que la normalización basada en modelar el crecimiento lograba un buen resultado visual, sus índices de clasificación son en general inferiores a los del resto. Esto puede ser debido a que, a pesar de haber tratado de eliminar el efecto que producía el hecho de que no todas las muestras tenían su máximo en la misma posición mediante un modelo relativo, el algoritmo de centrado (Capítulo 2) falla con algunas muestras cuya deriva del máximo se produce entre las longitudes de onda sobre las que se centra. Esto se puede tratar de reparar buscando alternativas para el centrado que generalicen mejor.

Observando las matrices de confusión de la Tabla 6.6 hay dos detalles destacables. El primero es que a pesar de que SOM tiene el índice más bajo, todo su error se concentra en un mismo punto: muestras que pertenecen a Iso pero que son clasificadas como Amin. Sin embargo GCS tiene menos error global pero está disperso. El otro es que el error de SOM es el dual al de 2NN. A diferencia de SOM, 2NN concentra su error en algunas muestras de Amin que son clasificadas como Iso. Este hecho hace que estos dos clasificadores potencialmente pueden ser utilizados juntos mediante alguna técnica de aprendizaje conjunto (*Ensemble Learning*) (Polikar 2006) para reducir el error, o por lo menos aumentar el nivel de confianza de la clasificación.

En la tabla 6.3 se muestran los resultados de clasificación para distintos órdenes de las derivadas de los espectros como datos de entrada. La tendencia general de los resultados es a empeorar con el orden, debido probablemente a la mayor dispersión de las curvas que existe a medidas que se realizan sucesivas derivadas. En las matrices de confusión de 2NN de la tabla 6.7

se observa que tanto para orden 1 y 2 se repite el error que ya existía usando los espectros sin derivar pero además aparecen otros.

En la tabla 6.4 aparecen los resultados cuando se reduce la dimensión de los datos con la selección del algoritmo genético (se utilizaron las longitudes de onda 651, 666, 677, 688 y 694 nm) y la extracción de PCA (Capítulo 3). Aquí encontramos el mejor resultado de todos, la clasificación sin errores de 2NN usando la selección de variables, como ya se vio en el citado capítulo. Quizás una búsqueda con el algoritmo genético usando como función de aptitud a SOM o a GCS optimizaría los resultados de estos, pero fue más óptimo utilizar 2NN teniendo en cuenta los tiempos de ejecución. Observando las matrices de confusión asociadas (Tabla 6.8), vemos que los errores están más dispersos que cuando se utiliza todo el espectro. Esto ocurre tanto para GCS y SOM con la selección de variables, como para 2NN con PCA.

De la tabla 6.5 se desprende que si tomamos los mejores resultados parciales de GCS en el proceso de crecimiento de la red, éstos se acercan mucho a los de 2NN con la ventaja de utilizarse tan solo alrededor de 30 vectores para comparar con las muestras de entrada.

	Max-Min	Crecimiento	Media y Varianza	Wavelet	Derivada Orden 1	Selección	Extracción
2NN	0.991	**1**	**1**	**1**	**1**	**1**	0.991
SOM	0.947	0.895	0.895	0.912	0.886	**0.990**	0.947
GCS	0.966	0.977	0.955	**0.989**	0.977	0.977	0.977

Tabla 6.10. Tabla de Resultados del test de clasificación

En la tabla 6.10 figuran los resultados de entrenar al clasificador con todas las muestras que anteriormente se habían dividido en distintos grupos para hacer la validación cruzada, y de clasificar las muestras de test. Vemos que nuevamente k-vecinos ofrece los mejores resultados, mejorando los obtenidos con el set de validación. Esto se debe a que el buen funcionamiento de k-vecinos depende de tener disponible un set de datos representativos de las distintas clases, y en este último caso es en el que contaba con mayor información. Vemos que por el contrario los resultados de SOM decaen mientras que los de GCS son más estables. Esto puede dar a entender que SOM tiene dificultad para generalizar cuando se le presentan demasiados datos, mientras que

a naturaleza adaptativa de GCS le puede ayudar a sufrir el mismo problema. En cuanto a la utilización de las distintas normalizaciones y formato de los datos en general estos últimos resultados siguen la tónica de los vistos con el set de validación. Destacar que el mejor resultado de SOM con diferencia es el que obtuvo usando las 5 bandas seleccionadas. Sorprende también que para este mismo clasificador la normalización con mejores resultados haya sido la del máximo y mínimo a diferencia de los otros dos clasificadores.

Resumen

En este capítulo se ha introducido en primer lugar el campo de estudio del aprendizaje automático, algunas de cuyas técnicas se han utilizado para el propósito de clasificar los espectros de emisión de fluorescencia del fitoplancton. Las técnicas descritas han sido k-vecinos, *self-organizing maps y growing cell structures*. En el caso de SOM se ha presentado una modificación de su código para dar cabida a una medida de distancia basada en ángulos. Por último se han presentado los resultados de clasificación obtenidos para distintas configuraciones de los datos.

7. Conclusiones y trabajos futuros

7.1 Conclusiones

Teniendo en cuenta los resultados obtenidos se puede afirmar que el espectro de emisión de fluorescencia por sí solo proporciona suficiente información discriminatoria para distinguir entre clases de fitoplancton, e incluso al nivel de género (categoría taxonómica), dado que Iso y Pl comparten la misma clase biológica. Esta afirmación tiene las siguientes condiciones: Por un lado esto es cierto para datos obtenidos de cultivos puros, es decir muestras de agua en las que solo hay presentes células de una misma especie. Segundo, el número de especies utilizado es relativamente bajo. Dada la alta similitud aparente de los espectros es de suponer que, aunque utilizando seis tipos de especies diferentes la precisión de la clasificación esté por encima del 95% en la mayoría de los casos, si aumentáramos su número habría mayores probabilidades de que varios de ellos tuvieran un importante solape de sus espectros.

En cuanto a las técnicas de clasificación, no es extraño que el algoritmo de k-vecinos pueda ofrecer tan buenos resultados si tenemos en cuenta que a pesar de que aparentemente los espectros utilizados tienen del orden de 100 puntos, se ha demostrado que la dimensión real de los datos es mucho menor. Con cinco o menos variables se puede conservar prácticamente toda la información. Esto da pie a discutir la necesidad o no de utilizar instrumentación basada en sensores hiperespectrales.

En aplicaciones que requieren la realización de muchas medidas en poco tiempo (p.ej un perfilador automático que durante su descenso o ascenso se desea que mida la variabilidad de los organismos presentes en la columna en una escala temporal muy pequeña) es evidente la ventaja que proporciona utilizar el espectro de emisión para cuya obtención solo hay que excitar a una λ. Sin embargo, puede que excitar a 5 longitudes de onda distintos (si tomamos como ejemplo la selección utilizada con el algoritmo genético) no suponga perder excesiva resolución espacial, mientras que el coste del instrumental necesario sería menor.

También se puede argumentar que el hecho de que el algoritmo genético haya encontrado cinco bandas tan concretas que permitan incluso mejorar la precisión en la clasificación, ha sido gracias a que se disponía de un espectro con una buena resolución (1 nm) y que, si se hubiera

aplicado sobre otras seis especies distintas puede que las bandas seleccionadas fueran otras. En ese caso la selección sería dependiente del problema en particular (especies que se desena distinguir) y sería necesario contar con espectros completos, por lo menos para aprender de ellos.

El caso de la extracción de variables es distinto en cuanto que, por definición, se necesitan todas las variables originales para poder aplicarlo. La utilidad de esta técnica es más la de mejorar los resultados de clasificadores que sean muy sensibles a la dimensión de los datos y no la de permitir reducir el coste de la adquisición de los datos.

Como ya se ha comentado los datos espectrales no parecen la mejor manera de estudiar el efecto de la dimensionalidad sobre los clasificadores debido a que la dimensión efectiva es mucho menor, pero sí podemos hacer una comparación general entre SOM y GCS. En base a los resultados obtenidos parece que la capacidad de GCS para adaptarse a los datos disponibles le permite, en general, ofrecer mejores resultados que SOM, incluso cuando solo se tiene en cuenta el resultado para el tamaño de red final. Además, como ya se comentó en el capítulo anterior, SOM parece tener dificultades cuando el set de aprendizaje cuenta con muchas muestras.

Volviendo un poco más atrás, al tema de similitud, destacar la utilidad de un método simple basado en librería de espectros para aprender más sobre los datos y saber qué nos espera cuando nos lancemos a clasificar. En cuanto a las distancias probadas, si los datos están normalizados, la distancia euclidiana a demostrado ser un indicativo fiable de lo diferentes que son dos especies, y dado que muchas técnicas se basan en ella, es una opción razonable.

De las dos transformaciones de los datos vistas, la de wavelet ha demostrado ser de gran utilidad teniendo en cuenta que ha servido para suavizar los datos y para normalizarlos mediante técnicas que han demostrado ser mejores que el resto en base a criterios objetivos (pruebas estadísticas e índices de clasificación). La derivada de los espectros contienen casi tanto poder discriminativo como las muestras originales pero no se puede afirmar que aporten nueva información puesto que los errores de clasificación tenían lugar entre las mismas clases. Conviene verificar sin embargo, si son exactamente las mismas muestras las que provocan los fallos de clasificación.

Por último se ha demostrado la importancia de utilizar una técnica de suavizado adaptada a los datos para evitar distorsionar sus propiedades, entre ellas las estadísticas, que pudieran ser de utilidad para métodos posteriores. La normalización utilizada también ha tenido en influencia en

los resultados, pero en este caso algunas técnicas simples como la de máximo y mínimo o la de media y varianza han demostrado ser suficientemente buenos en muchos de los casos.

7.2 Trabajos futuros

Como ya se ha comentado, el trabajo realizado en este proyecto se ha basado en unos datos de fluorescencia adquiridos sobre cultivos puros, donde sólo existía una especie. Desgraciadamente en el mar rara vez es posible encontrar esta situación tan ventajosa. Diferentes especies coexisten en el mismo medio y evidentemente esto plantea muchas dificultades para aplicar las técnicas empleadas en este PFC directamente en un entorno natural.

En primer lugar hay que analizar la forma en la que interactúa la fluorescencia de las distintas especies cuando es su superposición la que recibe el detector. Es necesario evaluar si lo que se recibe es una mezcla lineal o si por el contrario es una interacción más compleja de naturaleza no lineal. En el caso de que fuera lineal se podrían aplicar técnicas de linear unmixing para separar los espectros implicados. Algunos autores han optado por suponer que efectivamente la mezcla es lineal (Xupeng et al. 2010b) y aplican estos métodos.

Los parámetros que se desearían conocer en estos casos son el número de especies con presencia importante, la proporción de cada uno de ellos y de qué especies se trata. Para estudiar este escenario hay que realizar tomas de muestras en el laboratorio sobre cultivos con mezclas muy controladas para las que habría que utilizarse algún estimador fiable de la concentración de células de cada especie. Otra opción es hacer uso de modelos, aunque éstos pueden llegar a conclusiones equivocadas sobre la forma en la que la superposición se realiza.

Todo esto se resume en la necesidad de contar con datos, muchos datos tanto de laboratorio como del medio natural y a ser posible de un gran número de especies para evaluar hasta qué punto es posible extender los resultados si se utilizan estas otras clases. Esto también se aplica a la mejora en el diseño de los clasificadores que también requiere de mucha información para el diseño de sistemas que generalicen muy bien para ser efectivas en distintas condiciones y circunstancias.

Relacionado con esto último, los cultivos de los cuales provienen los datos utilizados fueron sometidos a unas condiciones muy concretas de luz y temperatura. El fitoplancton presente en el medio natural, según sea la profundidad donde se encuentre, la luminosidad y la temperatura

varían notablemente. Hay estudios (Neori et al. 1984) que revelan la fotoadaptación que realizan las células para adaptarse a entornos con condiciones tan dispares, modificando su composición pigmentaria y por tanto cambiando su respuesta a la luz. Son necesarios estudios que se centren en el efecto concreto que tienen estas circunstancias sobre el espectro de emisión de fluorescencia.

Respecto a los tipos de distancia, puede ser interesante probar alguna técnica de clasificación basada en la correlación dada la separabilidad relativa que parece proporcionar. Por otro lado el aprendizaje de distancia basado en las probabilidades bayesianas de las muestras podría ser utilizado por un clasificador jerárquico en forma de árbol para realizar comparaciones dos a dos y aprovechar la mayor separabilidad de las distancias ponderadas.

Por último, en este proyecto se seleccionaron algunas técnicas consideradas adecuadas para cada situación o problema que se trataba de resolver. La cantidad de técnicas disponibles relacionadas con el procesado de datos y con el aprendizaje automático aplicables a este problema no son abarcables por un solo proyecto. La experimentación con otras nuevas sin duda permitirán dar un paso más hacia la aplicación final en el mar.

8. Bibliografía

8.1 Referencias

Alpaydin, E. 'Introduction to machine learning' *2nd Ed The MIT Press, Cambridge*.

Anderson, D. M.; Yoshi, K. & White, A. W. (2000), 'Estimated annual economic impacts from harmful algal blooms (HABs) in the United States', *Woods Hole Oceanog. Inst. Tech. Rept.*, WHOI-2000-11.

Aymerich, I. F.; Piera, J.; Soria-Frisch, A. & Cros, L. (2009) 'A Rapid Technique for Classifying Phytoplankton Fluorescence Spectra Based on Self-Organizing Maps' *Appl. Spectrosc.,* vol. 63, pp. 716-726.

Chang, C. I. (2000), 'An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis', *IEEE Trans.Inf.Theory* 46(5), 1927--1932.

Chang, G.; Mahoney, K.; Briggs-Whitmire, A.; Kohler, D.; Mobley, C.; Lewis, M.; Moline, M.; Boss, E.; Kim, M.; Philpot, W. & T. Dickey (2004), 'The New Age of Hyperspectral Oceanography', *Oceanography,* vol. 17.

Chang, G.; Dickey, T. & Lewis, M. (2006), 'Toward a global ocean system for measurements of optical properties using remote sensing and in situ observations', *Remote Sensing of the Marine Environment: Manual of Remote Sensing, vol. 6, pp. 285-326*.

Cowles, T. J.; Desiderio, R. A. & Neuer, S. (1993), 'In situ characterization of phytoplankton from vertical profiles of fluorescence emission-spectra', *Mar.Biol.* 115(2), 217--222.

Donoho, D. L. (1995), 'De-noising by soft-thresholding', *IEEE Trans.Inf.Theory* 41(3), 613--627.

Donoho, D. L. & Johnstone, I. M. (1995), 'Adapting to unknown smoothness via wavelet shrinkage', *Journal of the American Statistical Association* 90(432), 1200--1224.

Du, P. J.; Fang, T.; Tang, H. & Shi, P. F. (2003), 'Encoding methods of spectral vector in hyperspectral remote sensing image', *Journal of Shanghai University*, Vol.9, No.1.

Franken, P.; Hill, A.; Peters, C. & Weinreich, G. (1961), 'Generation of Optical Harmonics', *Physical Review Letters* 7: 118.

Fritzke, B. (1993) 'Kohonen feature maps and growing cell structures – a performance comparison', *Advances in Neural Processing Systems 5*.

Fritzke, B. (1994), 'Growing Cell Structures - a Self-Organizing Network for Unsupervised and Supervised Learning', *Neural Networks* 7(9), 1441--1460.

García-Weil, L.; Arbelo, M. & Pérez-Marrero, J. (2009), 'La observación de los océanos desde el espacio', *in Oceanografía y Satélites, 1st ed.*, C. García-Soto, Ed. Madrid: Tébar, pp. 25-28.

Graps, A. (1995), 'An Introduction to Wavelets', *IEEE Comput.Sci.Eng.* 2(2), 50--61.

Kirk, J. T. O. 'Light and photosynthesis in aquatic ecosystems', *2nd ed. Cambridge University Press, Cambridge, 509* pp.

Kohonen, T. (1998), 'The self-organizing map', *Neurocomputing* 21(1-3), 1--6.

Kong, X.; Shu, N.; Huang, W. & Fu, J. (2010), 'The research on effectiveness of spectral similarity measures for hyperspectral image', *CISP2010*.

Lakowicz, J. R. 'Principles of Fluorescence Spectroscopy' *3rd Ed, Springer*.

Ledoit, O.; Wolf, M. (2002), 'Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size', *Annals of Statistics,* Vol 30, No 4, 1081-1102.

Levina, E. & Bickel, P.J. (2005). 'Maximum likelihood estimation of intrinsic dimension', In *Advances in NIPS* 17, Eds. L. K. Saul, Y. Weiss, L. Bottou.

Lewis, M. M. (2001), *Discriminating vegetation with hyperspectral imagery - What is possible?*.

Neori, A.; Holm-Hansen, O.; Mitchell, B. G. & Kiefer, D. A. (1984), 'Photoadaptation in marine phytoplankton', Plant Physiol, 76, 518-524.

Oldham, P. B.; Zillioux, E. J. & Warner, I. M. (1985), 'Spectral fingerprinting of phytoplankton populations by two-dimensional fluorescence and Fourier-Transform-Based pattern recognition', *J.Mar.Res.* 43(4), 893--906.

Perry, M. & Rudnick, D. (2003), 'Observing the ocean with autonomous and lagrangian platforms and sensors: the role of ALPS in sustained ocean observing systems," *Oceanography,* vol. 16.

Polikar, R. (2006), 'Ensemble based systems in decision making', *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21-45.

Pons, S.; Aymerich, I.; Torrecilla, E. & Piera, J. (2007) 'Monolithic spectrometer for environmental monitoring applications', *IEEE Oceans Europe Conference. Proceedings,* pp. 1-3.

Poryvkina, L.; Babichenko, S. & Leeben, Aina. (2000), 'Analysis of phytoplankton pigments by excitation spectra of fluorescence', *Proceedings of EARSeI-SIG-Workshop LIDAR.*

Randolph, T. W. (2006), 'Scale-based normalization of spectral data', *Cancer Biomarkers 2*, 135-144.

Robila, S. A. & Gershman, A. (2005), 'Spectral matching accuracy in processing hyperspectral data'.

Savitzky, A. & M. Golay, (1964), 'Smoothing and Differentiation of Data by Simplified Least Squares Procedures' *Analytical Chemistry*, 36(8):1627–1638.

Schofield, O.; Grzymski, J.; Bissett, W. P.; Kirkpatrick, G. J.; Millie, D. F.; Moline, M. & Roesler, C. S. (1999), 'Optical monitoring and forecasting systems for harmful algal blooms: Possibility or pipe dream?', *J.Phycol.* 35(6), 1477--1496.

Seppälä, J. (2009), 'Fluorescence properties of Baltic Sea phytoplankton' *Ph.D. Dissertation, Aquatic Sciences, Univ. of Helsinki, Finland*.

Taswell, C. (2000), 'The what, how, and why of wavelet shrinkage denoising', *Computing in Science & Engineering* 2(3), 12--19.

Torrecilla, E.; Piera, J. & Vilaseca, M. (2009), 'Derivative analysis of hyperspectral oceanographic data', *Advances in Geoscience and Remote Sensing, 1st ed., Gary Jedlovec, Ed. In-Tech*, 2009, pp. 597-618.

Tsai, F. & Philpot, W. (1998), 'Derivative analysis of hyperspectral data', *Remote Sens.Environ.* 66(1), 41--51.

Vaiphasa, C. (2006), 'Consideration of smoothing techniques for hyperspectral remote sensing', *Isprs Journal of Photogrammetry and Remote Sensing* 60(2), 91--99.

Vaiphasa, C.; Skidmore, A. K.; de Boer, W. F. & Vaiphasa, T. (2007), 'A hyperspectral band selector for plant species discrimination', *Isprs Journal of Photogrammetry and Remote Sensing* 62(3), 225--235.

Van Der Meer, F. & Bakker, W.(1997) 'CCSM: Cross correlogram spectral matching', *International Journal of Remote Sensing*, 18: 5, 1197 — 1201

Van der Meer, F. (2006), 'The effectiveness of spectral similarity measures for the analysis of hyperspectral imagery', *International Journal of Applied Earth Observation and Geoinformation* 8(1), 3--17.

Xing, E. P.; Ng, A. Y.; Jordan, M.I. & Russell, S. (2002), 'Distance Metric Learning, with application to Clustering with side-information', *Advances in Neural Information Processing Systems 16* (NIPS2002), *MIT Press,* 521-528.

Xupeng, H.; Rongguo, S.; Weiming, Z.; Shijun, R.; Hongtao, W.; Xiaoping, C. & Yiming, W. (2010a), 'Research on the discrimination methods of algae based on the fluorescence excitation spectra', *Acta Oceanologica Sinica* 29(4), 116--128.

Xupeng H..; Rongguo, S.; Fang, Z.; Xiulin, W.; Hongtao, W. & Zhixi, Z. (2010b), 'Multiple excitation wavelength fluorescence emission spectra technique for discrimination of phytoplankton.', *Journal of Ocean University of China Oceanic and Coastal Sea Research* 9(1), 16--24.

Yentsch, C. S. & Phinney, D. A. (1985), 'Spectral Fluorescence - an Ataxonomic Tool for Studying the Structure of Phytoplankton Populations', *J.Plankton Res.* 7(5), 617--632.

Zhang, Q. Q.; Lei, S. H.; Wang, X. L.; Wang, L. & Zhu, C. J. (2006), 'Discrimination of phytoplankton classes using characteristic spectra of 3D fluorescence spectra', *Spectrochimica Acta Part A-Molecular and Biomolecular Spectroscopy* 63(2), 361--369.

8.2 Bibliografía complementaria

Beutler, M.; Wiltshire, K. H.; Meyer, B.; Moldaenke, C.; Luring, C.; Meyerhofer, M.; Hansen, U. P. & Dau, H. (2002), 'A fluorometric method for the differentiation of algal populations in vivo and in situ', *Photosynthesis Res.* 72(1), 39--53.

Cowles, T. J. & Desiderio, R. A. (1993), 'Resolution of biological microstructure through in situ fluorescence emission spectra', *Oceanography*, Vol 6, No. 3.

(2004) 'Genetic algorithm and direct search toolbox user's guide'.

Ferrer. M. A.; Travieso, C. M. & Alonso, J. B. (2005) 'Tratamiento digital de la señal: Fundamentos y aplicaciones'.

Gan, F.; Hopke, P. K. & Wang, J. (2009), 'A spectral similarity measure using Bayesian statistics', *Anal.Chim.Acta* 635(2), 157--161.

Gregor, J. & Marsalek, B. (2005), 'A simple in vivo fluorescence method for the selective detection and quantification of freshwater cyanobacteria and eukaryotic algae', *Acta Hydrochim.Hydrobiol.* 33(2), 142--148.

Kotsiantis, S. B.; Zaharakis, I. D. & Pintelas, P. E. (2006), 'Machine learning: a review of classification and combining techniques', *Artif.Intell.Rev.* 26(3), 159--190.

Kotsiantis, S. B.; Kanellopoulos, D. & Pintelas, P. E. (2006), 'Data Preprocessing for Supervised Leaning', *Proceedings of World Academy of Science, Engineering and Technology, Vol 12* 12, 278--283.

Misiti, M.; Misiti, Y.; Oppenheim, G. & Poggi, J. M. (1996), 'Wavelet toolbox user's guide'.

Poryvkina, L.; Babichenko, S.; Kaitala, S.; Kuosa, H. & Shalapjonok, A. (1994), 'Spectral fluorescence signatures in the characterization of phytoplankton community composition', *J.Plankton Res.* 16(10), 1315--1327.

Ruffin, C.; King, R. L. & Younani, N. H. (2008), 'A combined derivative spectroscopy and Savitzky-Golay filtering method for the analysis of hyperspectral data', *Giscience & Remote Sensing* 45(1), 1--15.

Vesanto, J.; Himberg, J.; Alhoniemi, E. & Parhankangas, J. (2000), 'SOM toolbox for Matlab 5'.

Villmann, T.; Merényi, E. & Seiffert, U. (2008) 'Machine learning approaches and pattern recognition for spectral data', *16th European Symposium on Artificial Neural Networks.*